Ensemble Analysis for 2021 State Legislative Redistricting in Colorado

Jeanne Clelland[*], Daryl DeFord, Beth Malmskog, and Flavia Sancier-Barbosa

September 26, 2021

Abstract

In this report, we apply techniques of ensemble analysis to establish a baseline context for State Legislative redistricting in Colorado following the 2020 Census. We generate large random samples of redistricting plans for the State Senate and State House that meet the basic legal requirements established by Amendment Z. Using these samples, we establish "reasonable" ranges for what might be expected for minority population, competitive districts, and partisan seat share for plans generated without explicit consideration of these issues. We also explore how these various priorities interact; in particular, we explore how the constitutional imperative to keep counties whole as much as possible affects both the ability to maximize the number of competitive districts and the expected range for partisan seat share. Finally, we compare the First and Second Staff Plans for the State Senate and State House proposed by the Colorado Independent Legislative Redistricting Commission's nonpartisan staff to our ensembles and comment on their performance relative to the ensembles.

1 Introduction

In the years since the last decennial redistricting cycle, there has been much interest in—and litigation around—quantifying and identifying partisan bias in district plans. Unlike racial gerrymandering, which has historically been limited by the Voting Rights Act of 1965, partisan gerrymandering has largely been unchecked by the courts until fairly recently, primarily due to the difficulty of identifying a quantifiable standard for measuring it.

One recently developed strategy for quantifying partisan bias is the idea of "ensemble analysis," in which a particular district plan is compared to a large collection of randomly generated, legally valid plans, referred to as an "ensemble" of plans. This idea has been gaining traction in redistricting litigation in the last few years. For instance, Jonathan Mattingly, et. al. performed detailed

[*]The first author was partially supported by a Collaboration Grant for Mathematicians from the Simons Foundation.

ensemble analyses of North Carolina's Congressional [9] and state [10] legislative district plans that played key roles in the court cases [3] and [2], and Moon Duchin's ensemble analysis [8] of Pennsylvania's Congressional Districts played a similar role in [1]. Similar work can be found in Wesley Pegden's expert reports for Pennsylvania [11] and North Carolina [12].

The primary aim of our work is to use ensemble analysis to establish a baseline context for State Legislative redistricting in Colorado in 2021, in order to understand what might reasonably be expected for measures such as minority population, competitive districts, and partisan seat share, based on the state's unique political geography. This baseline may then be applied to evaluate proposed district plans under consideration by the Colorado Independent Congressional Legislative Commission to ensure that they satisfy the requirements specified by Amendment Z to the Colorado Constitution.

Here and throughout this report, we wish to emphasize that **none of the plans in our ensembles are intended for adoption.** Redistricting is fundamentally a human endeavor, and there are many important considerations that are difficult or impossible to fully incorporate into a computer-generated ensemble. The ensembles that we will discuss here are intended **only** to provide context to which proposed plans may be compared with regard to specific quantitative measures.

Additionally, we want to make the following points clear at the outset:

- The goal of ensemble analysis is **not** to identify a single "best" value for any measure (e.g., number of competitive districts, or numbers of seats won by each party), but rather to identify a **range** of values that would be reasonably likely for plans drawn without taking any partisan data into account. This analysis only raises concerns when a proposed plan is an **extreme outlier** relative to the range of values seen in an ensemble.

- Despite the ubiquity of descriptions such as, "this plan has X Democratic districts and Y Republican districts," this analysis does **not** predict future election outcomes or flag particular election outcomes as extreme. The election data used to evaluate plans for competitive districts and partisan seat share is based on past, **statewide** elections, whereas outcomes for future local, district-based elections may vary widely from those for statewide elections, depending on a variety of factors such as incumbent effects, candidate availability, and issues of particular local concern. Rather, the goal of this analysis is to identify **district plans** that may produce atypical outcomes across a variety of elections of different types.

- Because we cannot model all considerations that the Commission may take into account when drawing maps, plans that appear to be extreme outliers compared to an ensemble may in fact have perfectly reasonable explanations for their deviation from the ensemble. For example, drawing districts informed by the goals of satisfying the requirements of the Voting Rights Act, preserving communities of interest, and maximizing the number of competitive districts may affect plans in ways that are not well represented by our ensembles. In such cases, more information about the design criteria may be required in order to evaluate a plan on its merits.

2 Introduction to ensemble analysis

In this section we give a brief description of the main ideas and aims of ensemble analysis. For a more detailed treatment of our approach and methodology, please see our paper [5] and Appendix A.

The fundamental goal of ensemble analysis is to model the political geography of a region (in this case, the state of Colorado) in order to better understand what might be expected for a "typical" district plan for the state. Plans may be evaluated with regard to a variety of measures: partisan balance of election results, geographic compactness of districts, competitiveness of district elections, preservation of communities of interest, racial/ethnic population within districts, etc. The main idea is to create a large number of randomly generated, valid plans that satisfy all relevant legal constraints—an "ensemble" of plans. Measures of interest are then computed for each plan in the ensemble using real population and voting data. The result is a statistical range of possible outcomes for each measure, to which any proposed plan may be compared. If a proposed plan appears to be an extreme outlier compared to the ensemble, this may suggest that factors not included in the ensemble design may have played an important role in the plan's construction. Such factors may be desirable (e.g., preservation of communities of interest) or not (e.g., partisan gerrymandering).

For this type of analysis, it is natural to build districts from voting precincts, as these are the smallest geographic units for which voting data is readily available. This is just one of many reasons why the plans in our ensemble are generally unsuitable for adoption; the final plans will almost certainly divide many precincts in order to achieve their aims.

Our construction of ensembles begins with a data-rich map of Colorado's voting precincts as of 2020. Details of our processes for data collection and construction of this map are described in Appendix A.1, and details of the algorithms used to build our ensembles are described in Appendix A.2. For this initial analysis, for each chamber (Senate and House) we constructed three ensembles of 2,000,000 random maps each, incorporating some of the most fundamental constitutional requirements:

- **Contiguity:** The algorithm used to generate district plans automatically guarantees district contiguity; see Appendix A.2 for more details.
- **Population equality:** We have required that all plans in our ensembles have a population deviation of 5% or less between the least- and most-populous districts, as required by Amendment Z.

- **Compactness:** The algorithm used to generate district plans is designed to preferentially sample from more compact plans, and a large body of experimental evidence indicates that it is generally very effective in this endeavor. (See, e.g., [6].) No specific metric for measuring compactness is prescribed by Amendment Z, and we did not explicitly track any quantitative measure of district compactness. However, we have included a few of our randomly generated maps in Appendix A.2 to illustrate that their districts are generally reasonably compact.

- **Preservation of political subdivisions and communities of interest:** Our first ensemble, which we shall refer to as "county-neutral," did not incorporate any information regarding political subdivisions such as cities or counties or other communities of interest. Our second ensemble, which we shall refer to as "county-aware," added an algorithm described in Appendix A.2 to minimize the number of county splits. Our third ensemble, which we shall refer to as "tailored county-aware," used the county-aware algorithm and incorporated two additional constraints, based on input from the Commission:

  1. Plans in this ensemble never split any of the 27 counties with a 2020 Census population of 10,000 or less.

  2. Plans in this ensemble never split four communities of interest identified by the Commission:

     – the counties of Sedgwick, Phillips, Logan, Morgan, Washington, and Yuma in Northeast Colorado;

     – the counties of Saguache, Alamosa, Rio Grande, Conejos, Costilla, and Mineral in the San Luis Valley;

     – the counties of Archuleta, LaPlata, San Juan, and as much of Montezuma as possible, keeping the individual tribes whole;

     – the Roaring Fork Valley, including the communities of Aspen, Basalt, El Jebel, Carbondale, Glenwood Springs, Rifle, Silt, and Parachute.

  Comparing statistics across these three ensembles will help to quantify how prioritizing the preservation of these subdivisions and communities of interest affects other priorities, such as the ability to draw competitive districts.

In Sections 3 and 5, we will explore how our county-neutral, county-aware, and tailored county-aware ensembles of plans for the State Senate and State House, respectively, typically perform on the measures of county splits, minority representation, competitive districts, and partisan seat share. For the latter two metrics, we will focus on the composite "election" obtained by averaging partisan outcomes for the 8 statewide elections between 2016 and 2020 that have been identified by the Commission, specifically:

- the elections for President and U.S. Senator in 2016;
- the elections for Attorney General, Governor, Regent At Large, Secretary of State, and Treasurer in 2018;
- the election for U.S. Senator in 2020.

Then in Sections 4 and 6, we will provide detailed comparisons of the First and Second Staff Plans for each chamber to these ensembles.

Before embarking on separate analyses for each chamber, we will conclude this section with some general discussion about each of these metrics.

2.1 County splits

For each ensemble, we counted the numbers of county splits in each plan in two ways:

1. numbers of "counties split," which count the numbers of counties divided between more than one district;
2. numbers of "total county splits," which count the numbers of times counties are split.

So, e.g., if a county is divided between three districts, this counts as one split towards the "counties split" measure and two splits for the "total county splits" measure.

This measure will be primarily used to understand how plans in the county-neutral, county-aware, and tailored county-aware ensembles typically compare to human-drawn plans (as exemplified by the Staff Plans) regarding county splits. By computing other statistics of interest for all three ensembles, we hope to better understand how the choice to preserve counties and certain communities of interest affects other redistricting priorities.

2.2 Minority representation

After contiguity, population equality, and the Voting Rights Act, the next highest priority specified by Amendment Z (co-equal with district compactness and preservation of political subdivisions) is the preservation of communities of interest. This is perhaps the most difficult criterion to model algorithmically, as communities of interest vary widely in nature and in geographic extent, and many different types of communities of interest overlap in complicated ways. Even in our tailored county-aware ensemble, we were only able to take into account a few communities of interest with clearly defined geographic boundaries that the Commission identified as high-priority.

One very significant community of interest that does not have such clearly defined geographic boundaries is the minority population of the state. We will examine the proportions of (1) Hispanic voting age population, and (2) Non-White voting age population within each district in our ensembles. For context, we note that for the state as a whole, the Hispanic voting age population is approximately 19.2% of the total voting age population, and the Non-White voting age population is approximately 26.6% of the total voting age population. We have not received specific direction from the Commission regarding the creation of majority-minority districts, and so we have not attempted to incorporate any such criteria into our ensembles; however, there is still general agreement that districts should be drawn so as to give these communities adequate representation.

2.3 Competitive districts

Competitive districts are defined in Amendment Z as "having a reasonable potential for the party affiliation of the district's representative to change at least once between federal decennial censuses." The lack of a quantitative standard in this definition has led to much discussion regarding the adoption of a standard for determining which districts will be considered competitive, and the Commission has decided to base its measure of competitiveness on an average of partisan outcomes (based only on votes for Democratic and Republican candidates) from 8 statewide elections from 2016 through 2020:

- the elections for President and U.S. Senator in 2016;
- the elections for Attorney General, Governor, Regent At Large, Secretary of State, and Treasurer in 2018;
- the election for U.S. Senator in 2020.

Each of these elections is given equal weight, creating a "composite election" whose Democratic and Republican vote percentages in each district are equal to the averages of the Democratic and Republican vote percentages, respectively, for these 8 elections in that district.

A typical measure of competitiveness involves prescribing a "vote band" about the 50% mark, and any election whose Democratic and Republican vote shares fall within that band is considered competitive. The Commission has adopted an 8.5% vote band, so that any election for which the Democratic and Republican vote shares fall between 45.75% and 54.25% is considered competitive.

2.4 Partisan seat share

Partisan seat share—i.e., the number of seats won by each political party in a particular election—is not one of the considerations prescribed by Amendment Z for district plans, but it is perhaps the outcome that is of the greatest interest to the most people. We will conclude our discussion for each chamber with a description of ensemble statistics for this measure.

3 Ensemble statistics for the State Senate

The goal of this section is to describe the main statistical properties of our county-neutral, county-aware, and tailored county-aware ensembles in order to establish context for what might reasonably be expected for State Senate district plans in Colorado. In Section 4, we will provide a detailed comparison of the First and Second Staff Plans to these ensembles.

3.1 County splits

The histograms in Figure 1 describe what percentage of plans in each ensemble exhibited each value for the number of counties split and the number of total county splits over the observed ranges.

- For the county-neutral ensemble, the mean number of counties split was 31.3 and the mean number of total splits was 94.3.
- For the county-aware ensemble, the mean number of counties split was 18.7 and the mean number of total splits was 55.7.
- For the tailored county-aware ensemble, the mean number of counties split was 17.7 and the mean number of total splits was 40.8.

We note that even for the tailored county-aware ensemble, our algorithm does not minimize the number of counties split quite as well as the First and Second Staff Plans, which split 11 and 13 counties, respectively. It more closely approximates the number of total county splits in the Staff Plans, with 37 and 43 total county splits, respectively.[1]

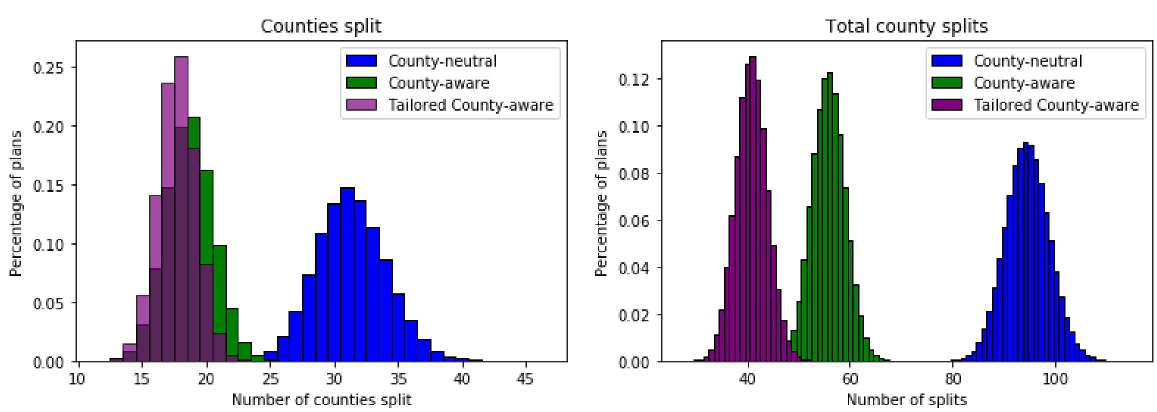


Figure 1: Counties split and total county splits for ensembles for State Senate

[1]These numbers were computed from approximations of the Staff Plans by districts made from whole precincts, and they may be slightly off from the true numbers.

3.2 Minority representation

For each plan in our ensembles, we computed the percentages of the Hispanic and Non-White voting age populations as a fraction of the total voting age population in each district and recorded the results. This data is displayed in Figures 2 and 3, organized as follows: For each plan, districts are sorted by Hispanic (resp., Non-White) voting age population percentage, from lowest to highest. The box plots show the ranges of these percentages for the sorted districts (in blue for the county-neutral ensemble, green for the county-aware ensemble, and purple for the tailored county-aware ensemble)—so, e.g., the second pair of boxes from the left shows the range of Hispanic (resp., Non-White) voting age population percentage in the second-lowest district in each plan. The boxes show the middle 50% of the range, and the whiskers extend from the 1st percentile through the 99th.

Additionally, because displaying all 35 districts in a single plot is somewhat unwieldy, we have also broken out the top 15 districts (highlighted with a green box in the full plot) into a separate plot below, with districts numbered from lowest to highest Hispanic (resp., Non-White) voting age population, using the numbers from the full sorted list of districts.

From these plots, we see that the preservation of counties and communities of interest has a very limited impact on the observed ranges for the minority populations of various districts. Furthermore, the observed ranges for districts with higher minority populations are fairly large, indicating that a wide range of values may occur in plans drawn without taking minority population into account.

3.3 Competitive districts

The histograms in Figure 4 describe what percentage of plans in each ensemble have each possible number of competitive districts according to the Commission's definition.

The mean numbers of competitive districts are 8.30, 8.13, and 8.98 for the county-neutral, county-aware, and tailored county-aware ensembles, respectively. Constraining county splits, at least within the ranges achieved by our ensembles, appears to have limited impact on the observed numbers of competitive districts, suggesting that it may not substantially affect the ability to draw competitive districts. However, since our ensembles did not achieve county split levels as low as the Staff Plans, we cannot say for certain whether this remains true for plans with fewer county splits than those in our ensembles.

For a more nuanced view on competitiveness, it is instructive to examine partisan outcomes by district. The box plots in Figure 5 are constructed similarly to those in Figures 2 and 3, except that now the boxes measure the observed ranges of Democratic vote share for each plan in the ensembles, ordered from most Republican to most Democratic. Also included in this plot are

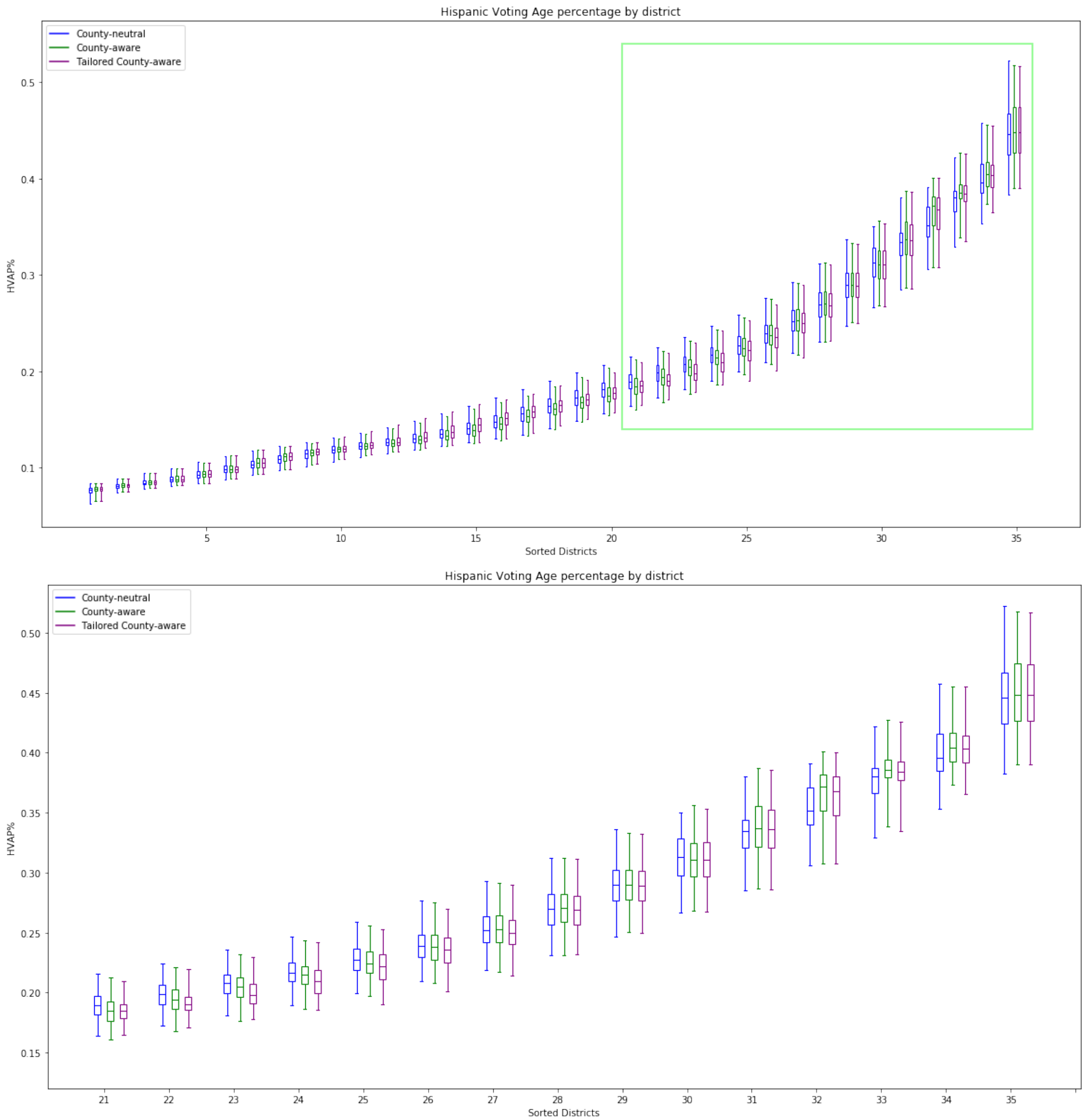


Figure 2: Hispanic voting age percentage by district for ensembles for State Senate

horizontal lines at the 50% mark and at the boundaries of the 8.5% vote band for reference.

From this figure we can make the following observations:

- The 8 most Republican and 13 most Democratic districts are essentially never competitive.
- The 9th most Republican district is competitive at the upper extreme of the county-neutral ensemble but very rarely in the county-aware and tailored county-aware ensembles, while the 14th most Democratic district is competitive at the lower extreme of the county-aware and

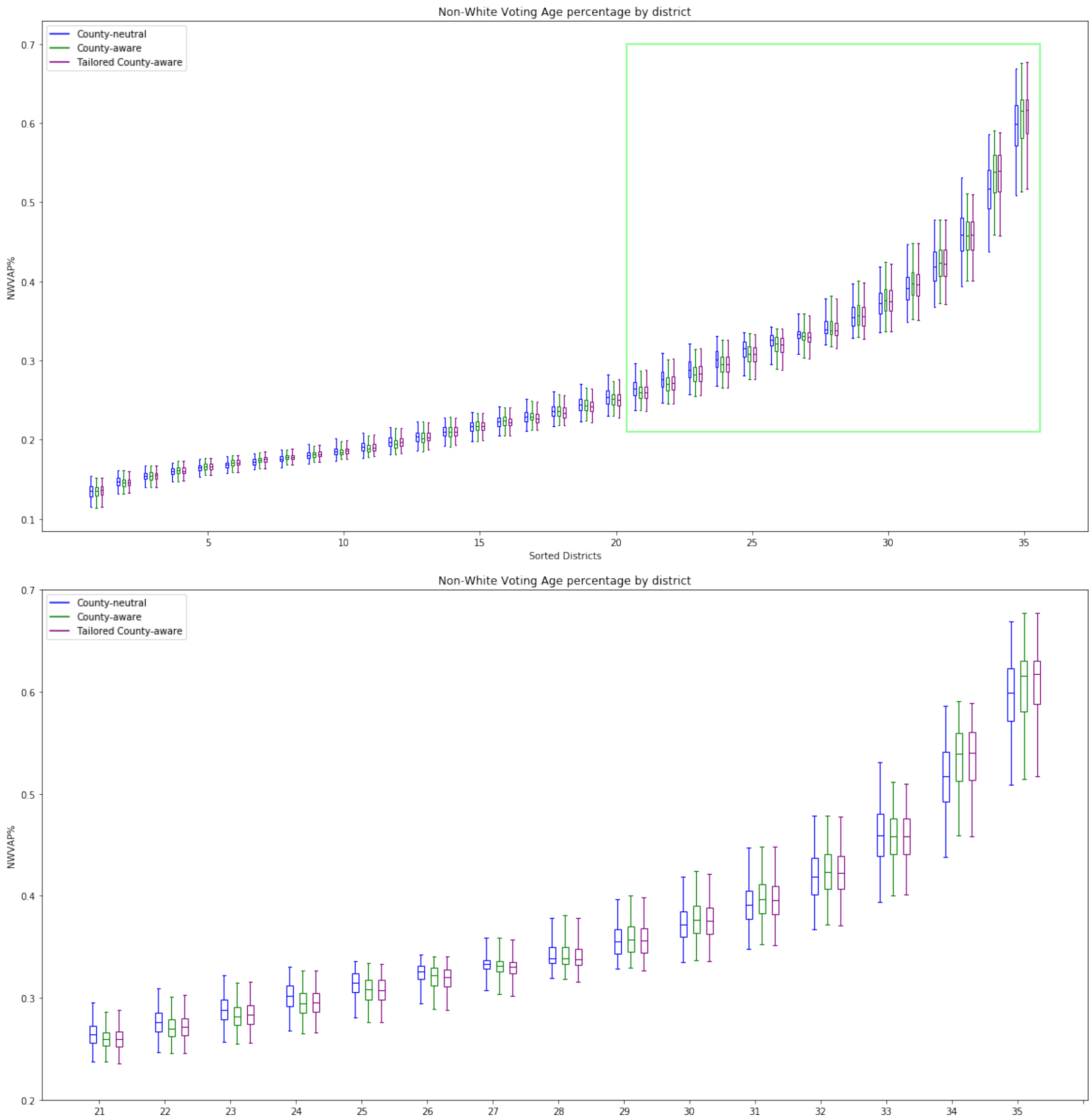


Figure 3: Non-White voting age percentage by district for ensembles for State Senate

tailored county-aware ensembles, but very rarely in the county-neutral ensemble.

- The 10th most Republican district and the 15th most Democratic district are occasionally competitive in all three ensembles.
- The 11th most Republican district is competitive about half the time in the county-neutral and tailored county-aware ensembles, and somewhat less often in the county-aware ensemble, while the 16th most Democratic district is competitive about half the time in the county-aware and tailored county-aware ensembles, and somewhat less often in the county-neutral

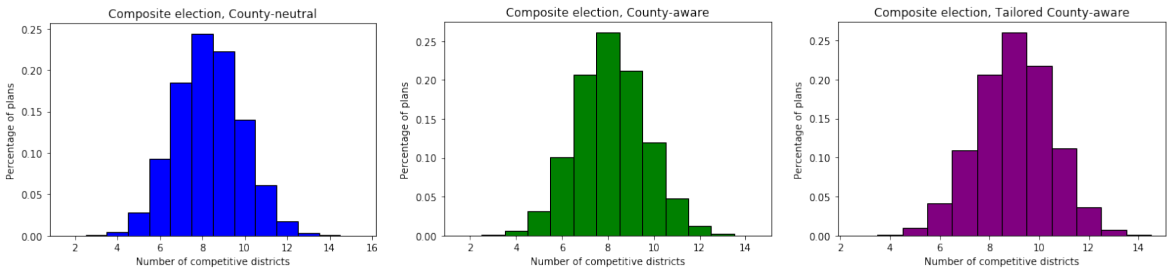


Figure 4: Numbers of competitive districts for ensembles for State Senate

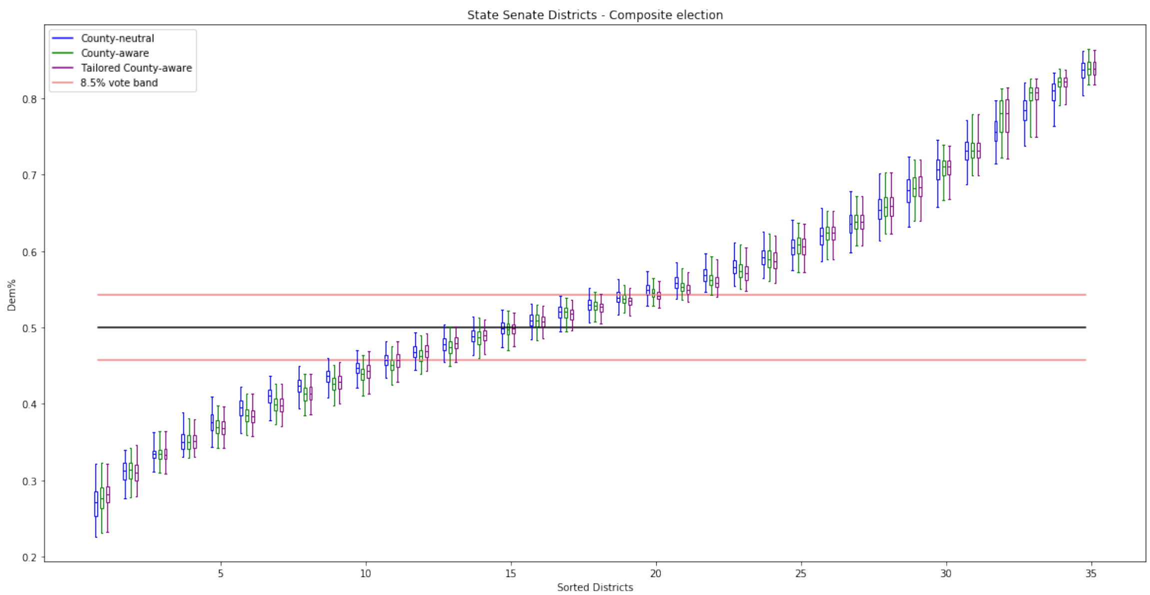


Figure 5: Democratic vote shares by district for ensembles for State Senate, with competitiveness vote band

ensemble.

- The 12th most Republican through 17th most Democratic districts (numbers 12 through 19 in the sorted list) are usually competitive in all three ensembles.

3.4 Partisan seat share

The histograms in Figure 6 describe what percentage of plans in each ensemble result in each possible number of Democratic seats won in the composite election. (The corresponding histograms for the numbers of Republican seats won would be the mirror images of the ones shown here.)

The most common outcomes in all three ensembles are 20 and 21 Democratic seats, with the latter outcome being slightly more common in the county-neutral ensemble than in the other

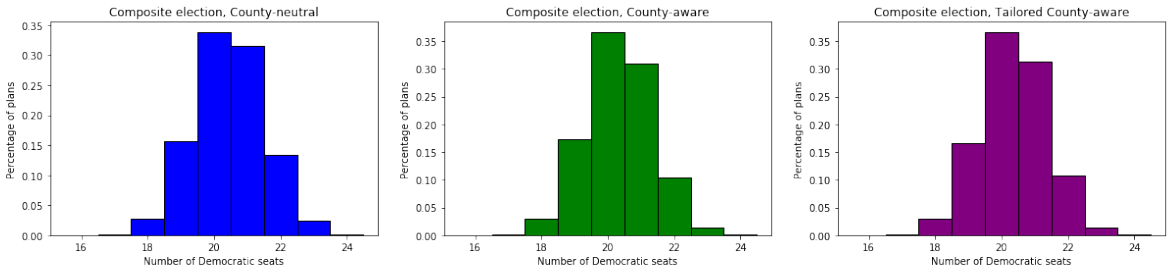


Figure 6: Numbers of Democratic seats won in composite election for ensembles for State Senate

two ensembles. The mean numbers of Democratic seats in the county-neutral, county-aware, and tailored county-aware ensembles are 20.45, 20.33, and 20.35, respectively.

As for competitive districts, we can see a more nuanced picture in the box plots of Figure 5. The districts numbered 13 through 17 in the sorted list might be considered "toss-up" districts, as their vote shares are, to varying degrees, reasonably likely to lie on either side of the 50% line. The variability of outcomes in these districts is responsible for the range of outcomes seen in the histograms in Figure 5.

4 Comparison of First and Second Staff Plans for State Senate to ensembles

On September 13, 2021, the Commission's nonpartisan staff released the First Staff Plan for State Senate districts, and on September 23, 2021, the Second Staff Plan was released. In this section we compare these plans to our ensembles for the measures described in the Section 3.

We wish to emphasize yet again that the Staff Plans are **absolutely not expected to be at or near the mean values** for either ensemble with respect to all the measures that we have computed. Even if the plans were drawn entirely randomly, about half of their computed values would be expected to lie outside the middle 50% range for the ensemble. Furthermore, the Commission and nonpartisan staff are not attempting to draw completely average plans, but rather to fulfill the Constitutional requirements that dictate that they attempt to preserve communities of interest and maximize the number of competitive districts. The comparison given here between the Staff Plans and our ensembles is intended **only** to provide context which may be used by the Commission as just one of many measures to evaluate the Staff Plans.

4.1 Minority representation

In Figures 7 and 8, we add the values for the Senate districts in the First and Second Staff Plans for the Hispanic voting age population and Non-White voting age population, respectively, to the box plots from Figures 2 and 3, showing only the top 20 districts in each case.

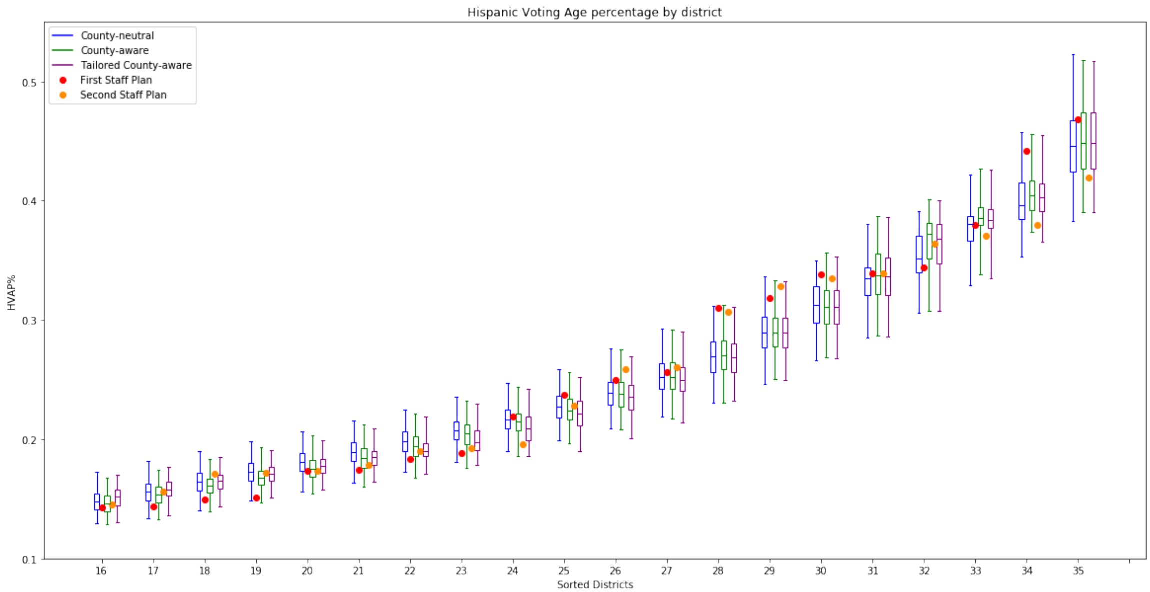


Figure 7: Hispanic voting age percentage by district for ensembles and Staff Plans for State Senate

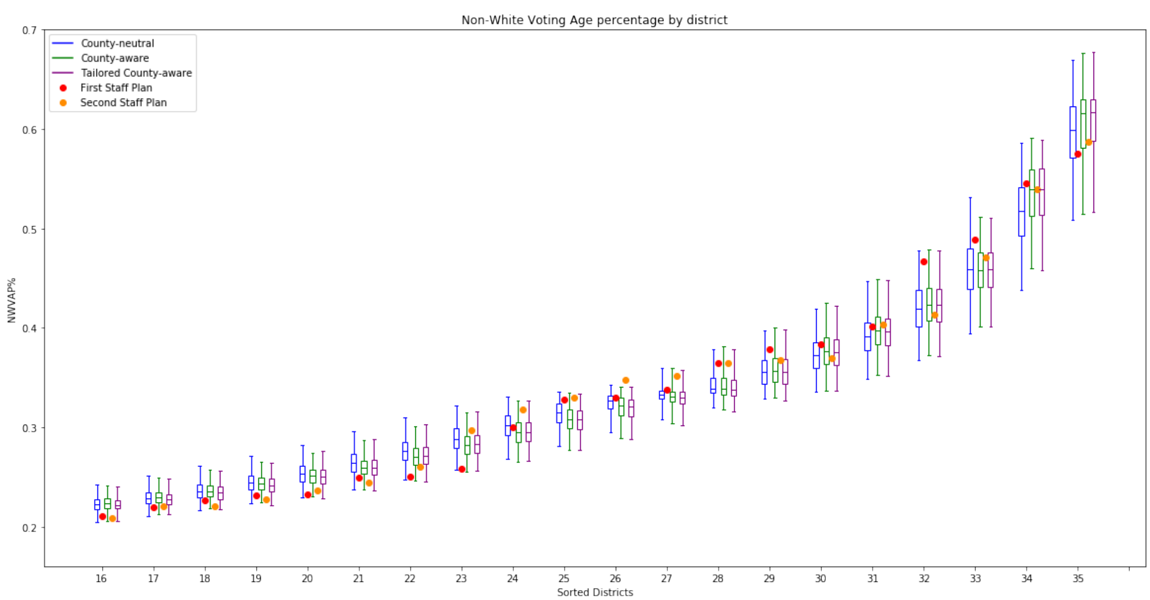


Figure 8: Non-White voting age percentage by district for ensembles and Staff Plans for State Senate

We do not see any extreme outliers with respect to either Hispanic or Non-White voting age populations. Both Staff Plans have 8 Senate districts with Hispanic voting age population above 30%, and the populations in the 7th and 8th highest districts are substantially above the means for all three ensembles, particularly in the Second Staff Plan. There is a substantial drop between the 8th and 9th highest districts, which suggests that the Staff Plans may have been deliberately designed to achieve 8 districts above this threshold. We also note that the two districts with the highest Hispanic voting age population have substantially higher percentages (46.8% and 44.2%) in the First Staff Plan than in the Second Staff Plan (41.9% and 38.0%).

Staff Plan 1 has 12 districts with Non-White voting age population above 29% and total Non-White population above 30%, and Staff Plan 2 has 13 such districts. The Non-White voting age population in the 13th highest district in the Second Staff Plan is slightly above the middle 50% for all three ensembles, but it is not an outlier. There is a substantial drop between the 13th and 14th highest districts in the Second Staff Plan (and between the 12th and 13th highest districts in the First Staff plan), which suggests that the Staff Plans may have been deliberately designed to achieve 12 and 13 districts, respectively, above this threshold.

4.2 Competitive districts

In Figure 9, we have added the values for the districts in the First and Second Staff Plans for the number of competitive districts to the histograms from Figure 4.

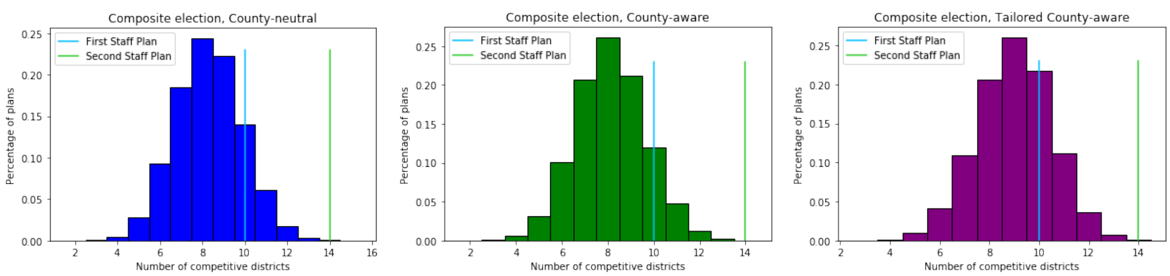


Figure 9: Numbers of competitive seats for ensembles and Staff Plans for State Senate

The First Staff Plan contains 10 competitive districts, which is slightly above the mean for all three ensembles. The Second Staff Plan contains 14 competitive districts, which is an extreme outlier for all three ensembles. However, since Amendment Z directs the Commission to maximize the number of competitive districts, this does not raise any concerns; rather, it indicates that this plan does an exceptional job of satisfying this constitutional priority.

In Figure 10, we have added the values for the districts in the First and Second Staff Plans to the box plots for the Democratic vote share for the composite election from Figure 5. We have also broken out the districts in the range from the 9th most Republican to the 14th most Democratic districts into a separate plot, as these are the districts with the potential to be competitive.

Here we can see clearly how the Second Staff Plan has improved upon the First Staff Plan by increasing the Democratic percentages in the most Republican districts in this range and increasing

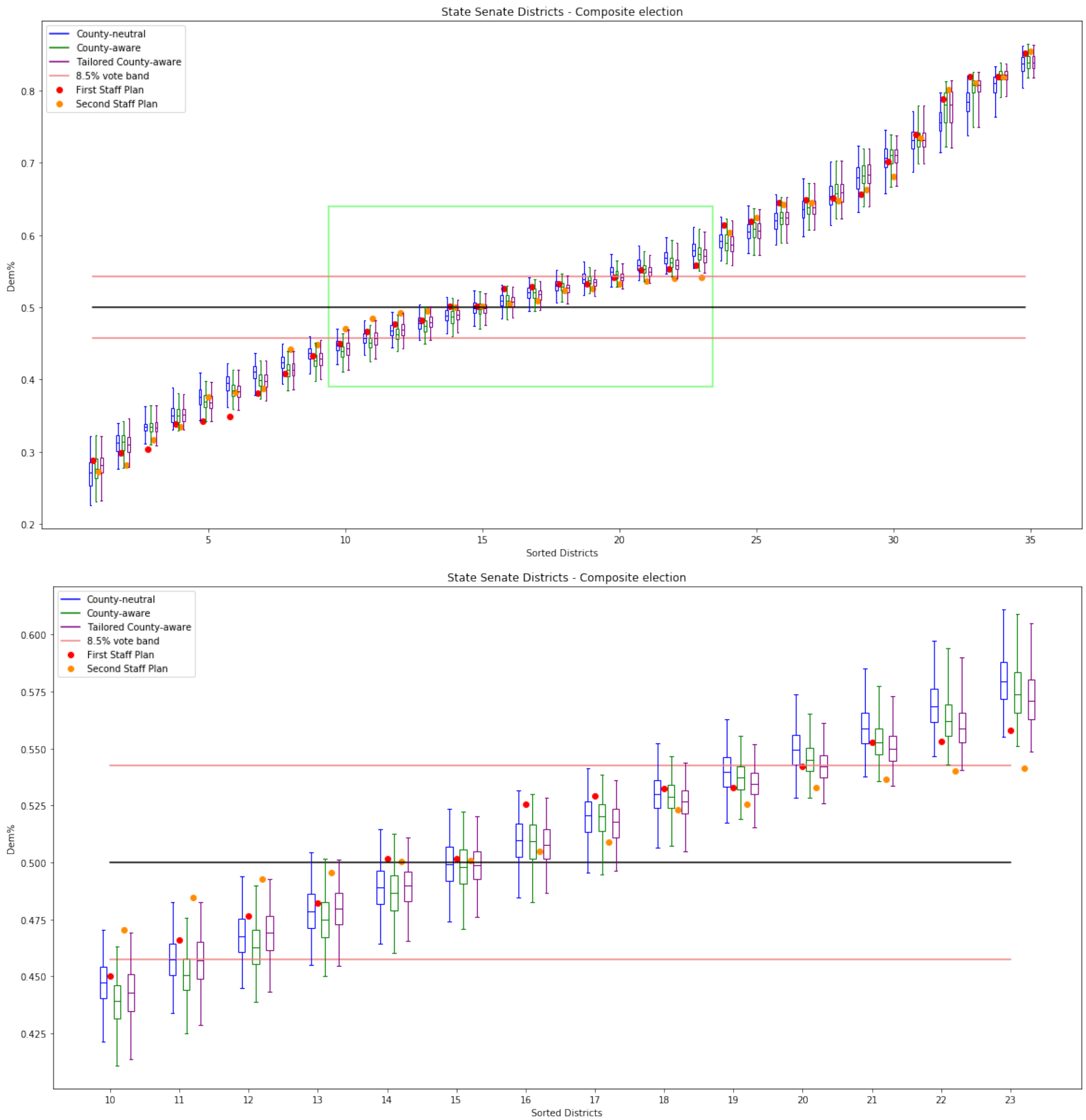


Figure 10: Democratic vote shares by district for ensembles and Staff Plans for State Senate, with competitiveness vote band

the Republican percentages in the most Democratic districts in this range, thereby creating some districts that are fairly extreme outliers relative to all three ensembles.

4.3 Partisan seat share

Finally, we compare the First and Second Staff Plans to our ensembles regarding partisan seat share. In Figure 11, we have added the values for the districts in both Staff Plans for the number of Democratic seats in the composite election to the histograms from Figure 6.

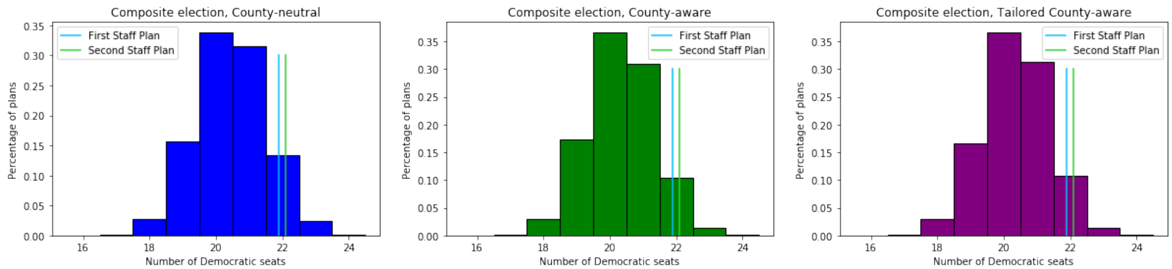


Figure 11: Numbers of Democratic seats won in ensembles and Staff Plans for State Senate

Both Staff Plans produce 22 Democratic seats, which is slightly above the mean for all ensembles, but still well within the range of reasonable outcomes. Moreover, as we can see from Figure 10, the bottom 2 Democratic seats are extremely competitive, with Democratic vote shares of 50.15% each in the First Staff Plan and Democratic vote shares of 50.05% and 50.10% in the Second Staff Plan. So these seats might reasonably be viewed as “toss-up” seats rather than either Democratic or Republican.

In order to explore this idea of “toss-up seats” further, we considered the range of partisan outcomes for the 8 statewide elections included in the composite election. Statewide Democratic vote shares for these elections ranged from 52.7% (President 2016) to 55.2% (Governor 2018), with an average for the composite election of 54.0%. This means that across these 8 elections, Democratic vote shares for the average district ranged from 1.3% below to 1.2% above the figure reported for the composite election. In particular, a typical district with reported Democratic vote share between 48.7% and 51.2% probably experienced majority votes for both parties at some point during these 8 elections.

With this in mind, we decided to explore an alternative classification of district-based partisan outcomes into three categories based on a 3% vote band about 50%:

1. Democratic: Democratic vote share of 51.5% or more;
2. Republican: Democratic vote share of 48.5% or less;
3. Toss-up: Democratic vote share between 48.5% and 51.5%.

The histograms in Figure 12 describe what percentage of plans in each ensemble fall into each of these three categories, with the values for the First and Second Staff Plans included for comparison.

These pictures tell an interesting story; the First Staff Plan is at or within one seat of the most common outcomes for all ensembles with respect to the numbers Democratic, Republican, and toss-up seats. The Second Staff Plan, on the other hand, has an extremely high number of toss-up seats, and consequently falls one or two seats below the ensemble means for both Democratic and

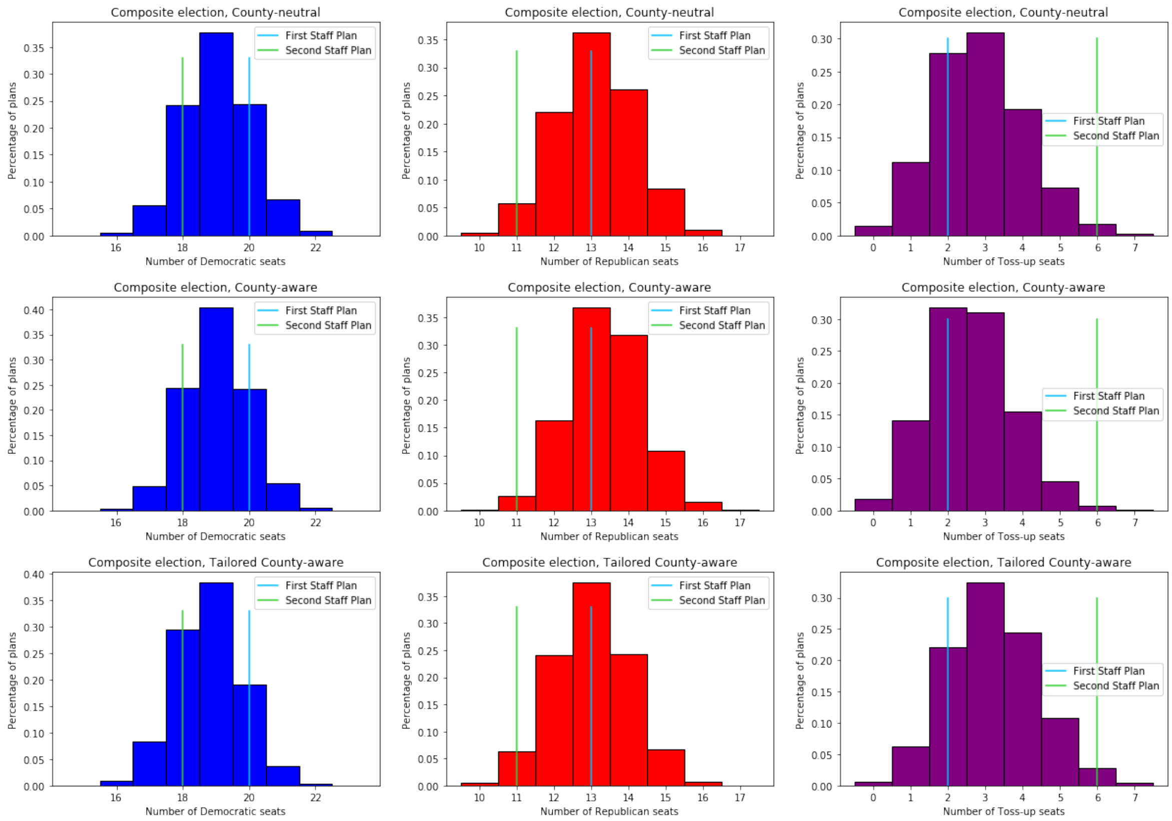


Figure 12: Numbers of Democratic, Republican, and Toss-up seats in ensembles and Staff Plans for State Senate

Republican seats. This indicates that the Second Staff Plan not only produces an unusual number of competitive seats within the 8.5% vote band chosen by the Commission, but it also produces an unusual number of **highly** competitive seats, as measured by the stricter 3% vote band that we have used to categorize toss-up seats.

5 Ensemble statistics for the State House

The goal of this section is to describe the main statistical properties of our county-neutral, county-aware, and tailored county-aware ensembles in order to establish context for what might reasonably be expected for State House district plans in Colorado. In Section 6, we will provide a detailed comparison of the First and Second Staff Plans to these ensembles.

5.1 County splits

The histograms in Figure 13 describe what percentage of plans in each ensemble exhibited each value for the number of counties split and the number of total county splits over the observed ranges. For the county-neutral ensemble, the mean number of counties split was 36.9 and the mean number of total splits was 140.3. For the county-aware ensemble, the mean number of counties split was 23.4 and the mean number of total splits was 91.1. For the tailored county-aware ensemble, the mean number of counties split was 20.0 and the mean number of total splits was 74.7.

By comparison, the First and Second Staff Plans, split 18 and 14 counties respectively, and they have 77 and 74 total county splits respectively.[2] This suggests that the tailored couunty-aware ensemble does a fairly reasonable job of sampling from plans that prioritize keeping counties whole to a similar degree as typical human-drawn plans.

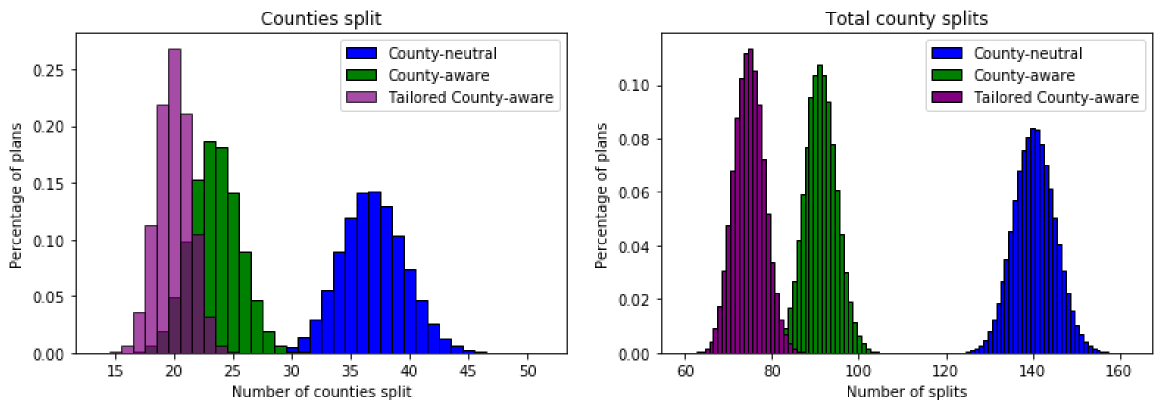


Figure 13: Counties split and total county splits for ensembles for State House

5.2 Minority representation

For each plan in our ensembles, we computed the percentages of the Hispanic and Non-White voting age populations as a fraction of the total voting age population in each district and recorded the results. This data is displayed in Figures 14 and 15, organized as follows: For each plan, districts are sorted by Hispanic (resp., Non-White) voting age population percentage, from lowest to highest. The box plots show the ranges of these percentages for the sorted districts (in blue for the county-neutral ensemble, green for the county-aware ensemble, and purple for the tailored county-aware ensemble)—so, e.g., the second pair of boxes from the left shows the range of Hispanic (resp., Non-White) voting age population percentage in the second-lowest district in each plan. The boxes show the middle 50% of the range, and the whiskers extend from the 1st percentile through the 99th.

Additionally, because displaying all 65 districts in a single plot is somewhat unwieldy, we have also broken out the top 20 districts (highlighted with a green box in the full plot) into a separate

[2]These numbers were computed from approximations of the Staff Plans by districts made from whole precincts, and they may be slightly off from the true numbers.

plot below, with districts numbered from lowest to highest Hispanic (resp., Non-White) voting age population, using the numbers from the full sorted list of districts.

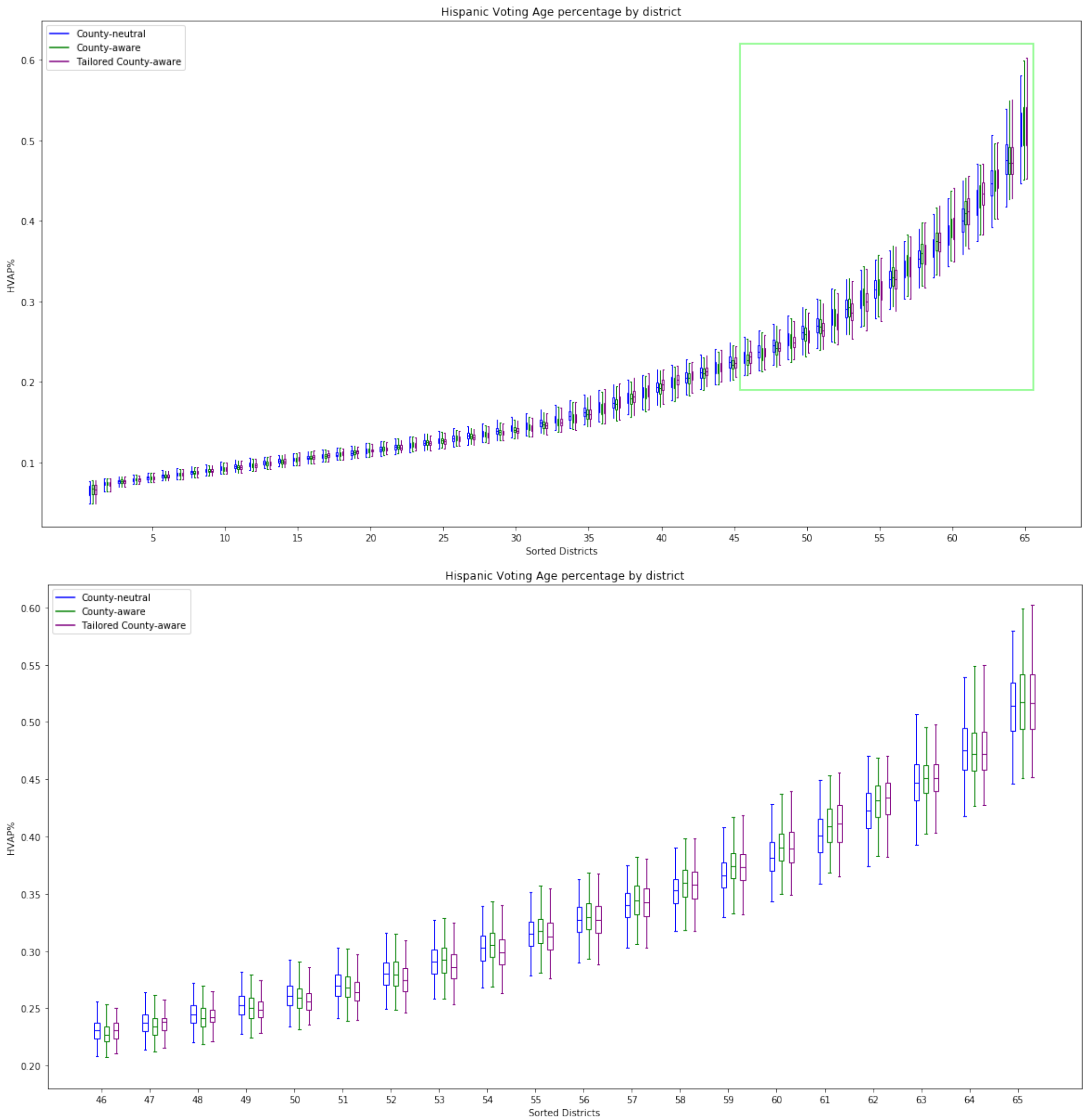
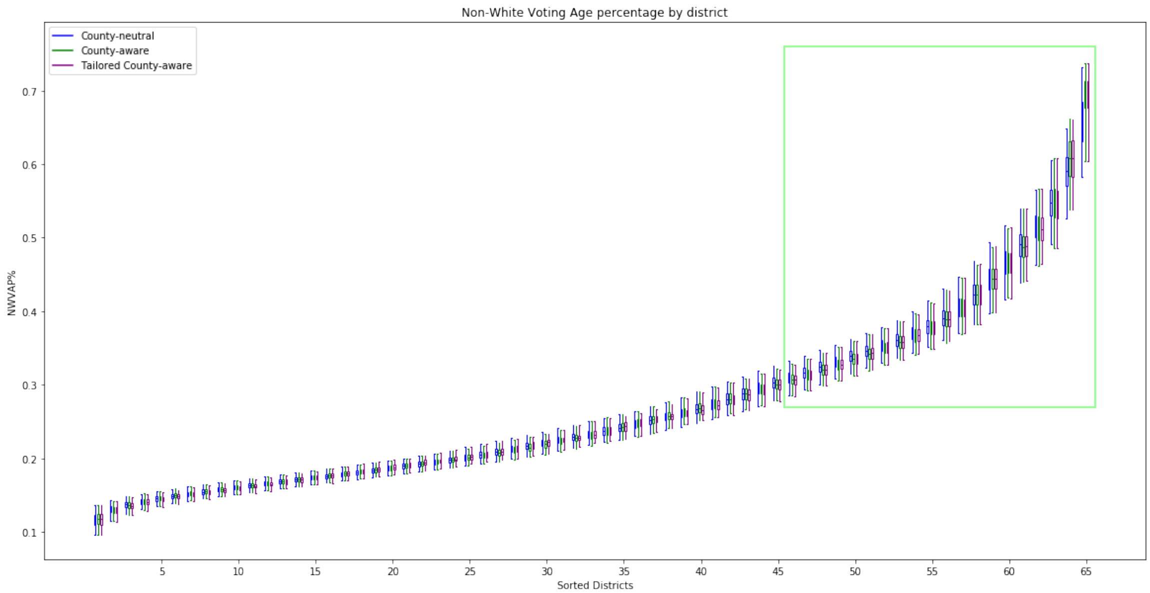


Figure 14: Hispanic voting age percentage by district for ensembles for State House

The only significant difference between the ensembles occurs for the district with the highest Non-White voting age population (and, to a lesser extent, for the second-highest district), where the county-aware and tailored county-aware ensembles tend to have higher percentages than the county-neutral ensemble. For both Hispanic and Non-White voting age populations, the observed ranges for districts with higher minority populations are fairly large, indicating that a wide range of values may occur in plans drawn without taking minority population into account.



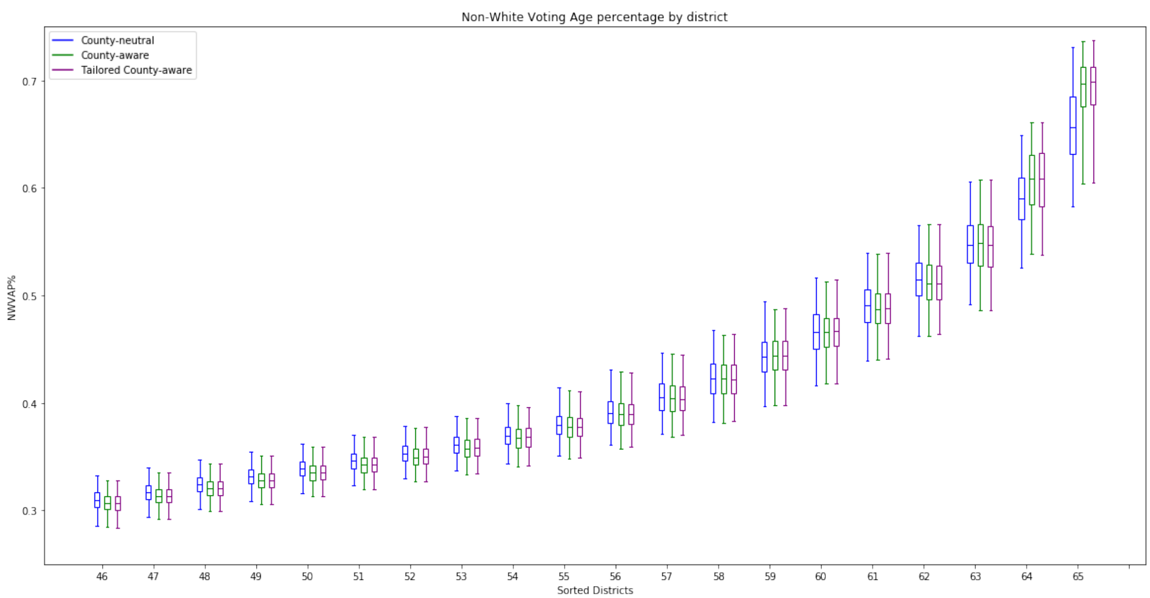


Figure 15: Non-White voting age percentage by district for ensembles for State House

5.3 Competitive districts

The histograms in Figure 16 describe what percentage of plans in each ensemble have each possible number of competitive districts according to the Commission's definition.

The mean numbers of competitive districts are 14.40, 13.93, and 13.63 for the county-neutral, county-aware, and tailored county-aware ensembles, respectively. Constraining county splits, at least within the ranges achieved by our ensembles, appears to have only a modest impact on the observed numbers of competitive districts, with plans with fewer county splits having slightly fewer

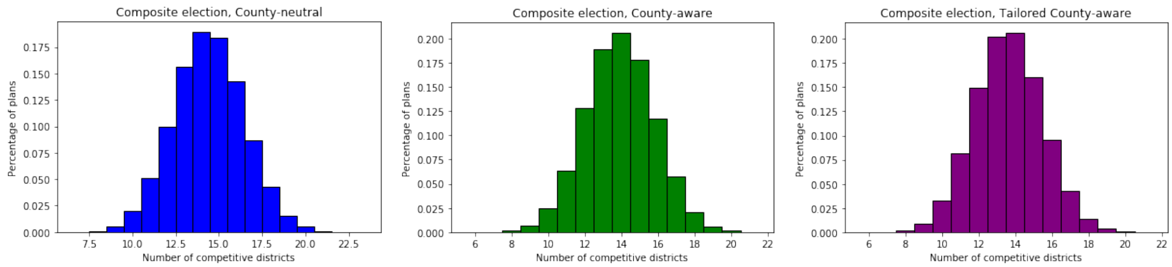


Figure 16: Numbers of competitive districts for ensembles for State House

competitive districts on average. Additionally, the range of reasonable values within each ensemble is much larger than this small difference between the ensemble means, so in practice this suggests that constraining county splits may not substantially affect the ability to draw competitive districts.

For a more nuanced view on competitiveness, it is instructive to examine partisan outcomes by district. The box plots in Figure 17 are constructed similarly to those in Figures 14 and 15, except that now the boxes measure the observed ranges of Democratic vote share for each plan in the ensembles, ordered from most Republican to most Democratic. Also included in this plot are horizontal lines at the 50% mark and at the boundaries of the 8.5% vote band for reference.

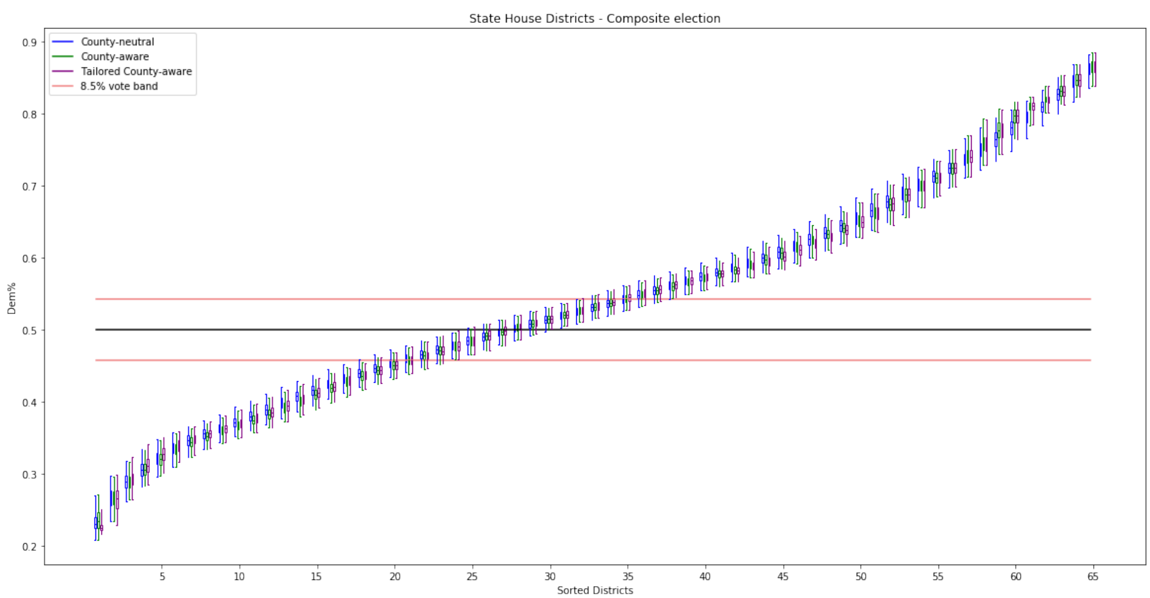


Figure 17: Democratic vote shares by district for ensembles for State House, with competitiveness vote band

From this figure we can make the following observations:

- The 17 most Republican and 27 most Democratic districts are essentially never competitive.

- The 18th most Republican district is competitive at the upper extreme of the county-neutral ensemble but very rarely in the county-aware and tailored county-aware ensembles, while the 28th most Democratic district is competitive at the lower extreme of the county-neutral and county-aware ensembles, but very rarely in the tailored county-aware ensemble.
- The 19th and 20th most Republican districts and the 29th and 30th most Democratic district are occasionally competitive in all three ensembles.
- The 21st most Republican district and the 31st most Democratic district are competitive about half the time in all three ensembles.
- The 22nd most Republican through 32nd most Democratic districts (numbers 22 through 34 in the sorted list) are usually competitive in all three ensembles.

5.4 Partisan seat share

The histograms in Figure 18 describe what percentage of plans in each ensemble result in each possible number of Democratic seats won in the composite election. (The corresponding histograms for the numbers of Republican seats won would be the mirror images of the ones shown here.)

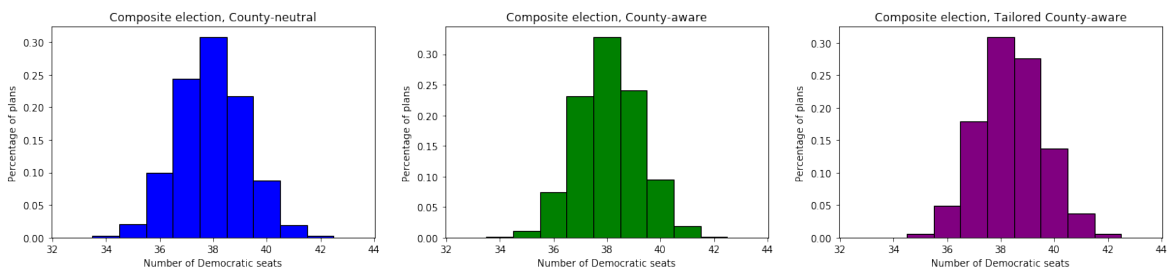


Figure 18: Numbers of Democratic seats won in composite election for ensembles for State House

The most common outcome in all three ensembles is 38 Democratic seats, with 37 and 39 seats also being relatively common. The mean numbers of Democratic seats in the county-neutral, county-aware, and tailored county-aware ensembles are 37.94, 38.08, and 38.39, respectively.

As for competitive districts, we can see a more nuanced picture in the box plots of Figure 17. The districts numbered 25 through 30 in the sorted list might be considered "toss-up" districts, as their vote shares are, to varying degrees, reasonably likely to lie on either side of the 50% line. The variability of outcomes in these districts is responsible for the range of outcomes seen in the histograms in Figure 17.

6 Comparison of First and Second Staff Plans for State House to ensembles

On September 13, 2021, the Commission's nonpartisan staff released the First Staff Plan for State House districts, and on September 23, 2021, the Second Staff Plan was released. In this section we compare these plans to our ensembles for the measures described in the Section 5.

We wish to emphasize yet again that the Staff Plans are **absolutely not expected to be at or near the mean values** for either ensemble with respect to all the measures that we have computed. Even if the plans were drawn entirely randomly, about half of their computed values would be expected to lie outside the middle 50% range for the ensemble. Furthermore, the Commission and nonpartisan staff are not attempting to draw completely average plans, but rather to fulfill the Constitutional requirements that dictate that they attempt to preserve communities of interest and maximize the number of competitive districts. The comparison given here between the Staff Plans and our ensembles is intended **only** to provide context which may be used by the Commission as just one of many measures to evaluate the Staff Plans.

6.1 Minority representation

In Figures 19 and 20, we add the values for the House districts in the First and Second Staff Plans for the Hispanic voting age population and Non-White voting age population, respectively, to the box plots from Figures 14 and 15, showing only the top 25 districts in each case.

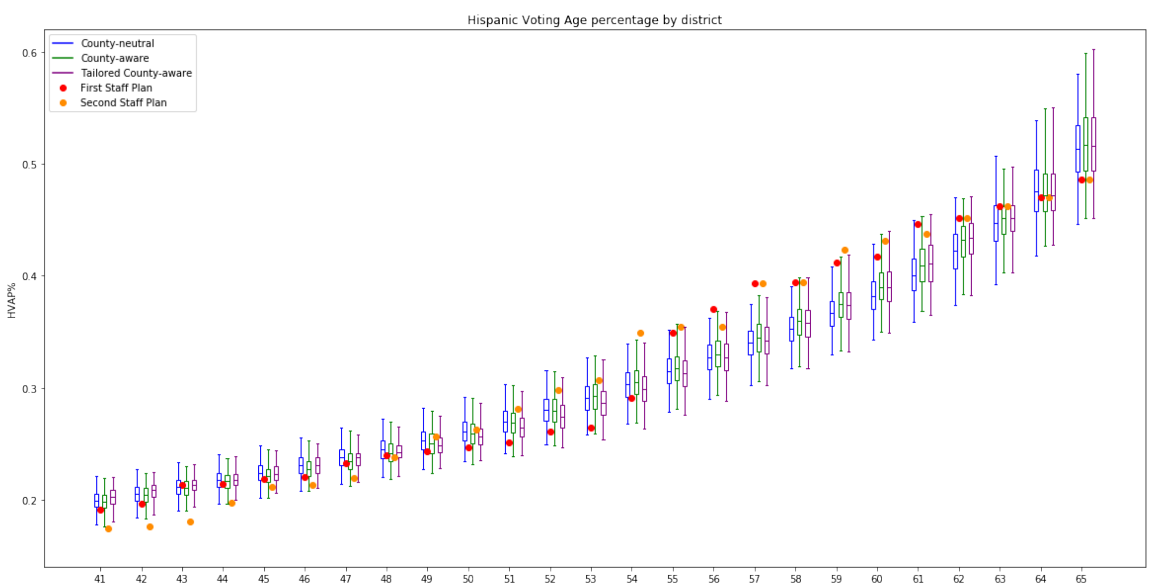


Figure 19: Hispanic voting age percentage by district for ensembles and Staff Plans for State House

The First Staff Plan has 11 districts with Hispanic voting age population above 30%, with the

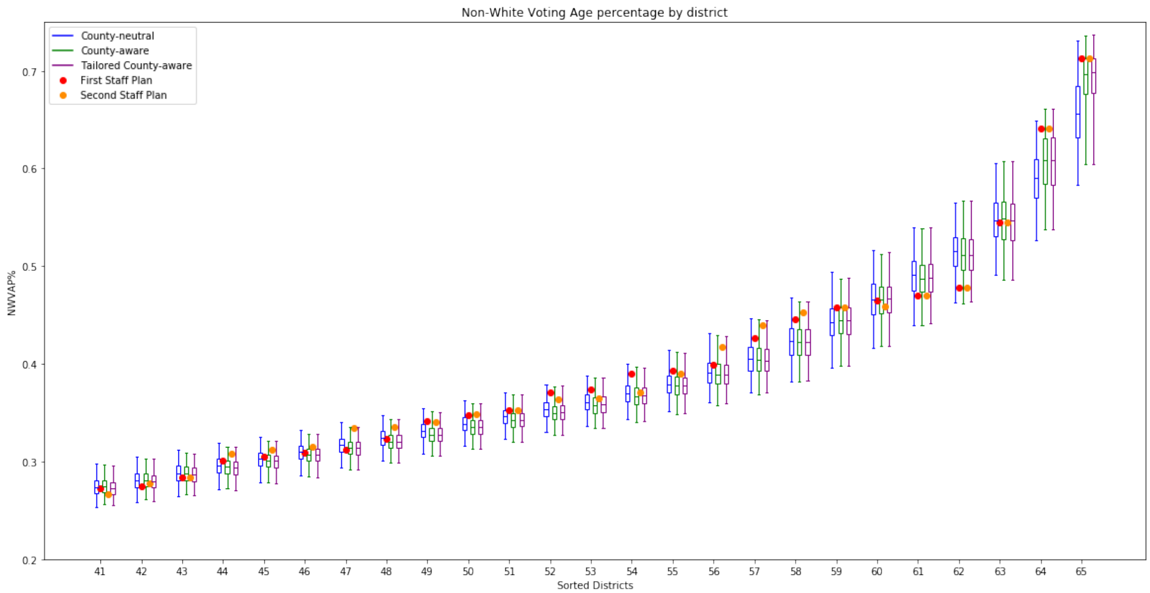


Figure 20: Non-White voting age percentage by district for ensembles and Staff Plans for State House

population of the 9th highest district particularly high relative to all three ensembles. The Second Staff Plan has 14 districts with Hispanic voting age population above 30%, with the populations of the 9th and 12th highest districts particularly high relative to all three ensembles. There is a substantial drop between the 11th and 12th highest districts in the First Staff Plan, and between the 12th and 13th highest districts in the Second Staff Plan, which suggests that the Staff Plans may have been deliberately designed to achieve 11 and 12 districts, respectively, above this threshold. (Although, as noted above, the Second Staff Plan actually contains two additional districts above the 30% threshold.)

Both Staff Plans have 22 districts with Non-White voting age population above 30%. The 22nd highest district has a slightly higher Non-White voting age percentage in the Second Staff Plan (30.8%) than in the First Staff Plan (30.1%), with the next-highest district in both plans having a Non-White voting age percentage of 28.4%.

6.2 Competitive districts

In Figure 21, we have added the values for the districts in the First and Second Staff Plans for the number of competitive districts to the histograms from Figure 16.

The First Staff Plan contains 11 competitive districts, including one that is exactly at the 8.5% competitiveness threshold. The Second Staff Plan contains 10 competitive districts, plus two additional districts within a 10% vote band about the 50% mark. These numbers are somewhat

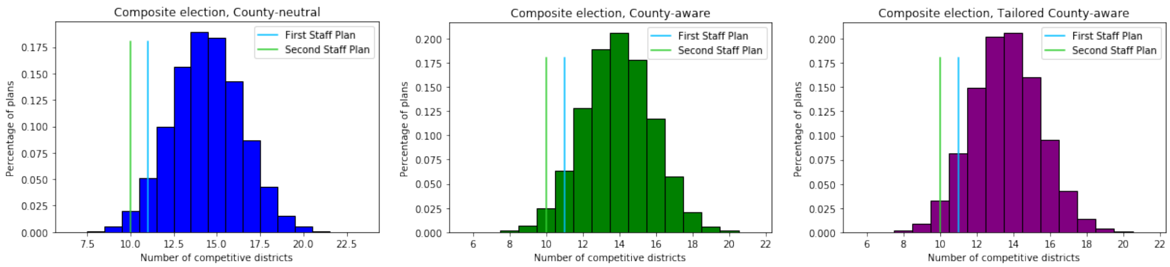


Figure 21: Numbers of competitive seats for ensembles and Staff Plans for State House

below the means for all three ensembles; they are not extreme outliers, but they do not provide evidence that the Staff Plans maximize the numbers of competitive districts. However, as we have repeatedly cautioned, they also do not necessarily indicate that the Staff Plans fail to satisfy this priority, as the discrepancy could easily result from prioritizing communities of interest or specific county or municipal boundaries that were not included in the design of our ensembles. For instance, the deliberate creation of more districts with Hispanic voting age percentage above 30% than expected, as seen in the Second Staff Plan, could potentially have an impact on the expected number of competitive districts.

With additional information from the Commission regarding what regions should be kept together in accordance with the Voting Rights Act, what county and/or municipal boundaries were intentionally prioritized, and what other communities of interest were preserved, we might be able to create ensembles that reflect the full range of constitutional criteria. This additional information is especially likely to influence the range of partisan outcomes and numbers of competitive districts seen in ensembles for State House plans, since State House districts consist of few enough people that keeping together relatively small geographic areas as communities of interest can greatly constrain House district boundaries within the region.

In Figure 22, we have added the values for the districts in the First and Second Staff Plans to the box plots for the Democratic vote share for the composite election from Figure 17. We have also broken out the districts in the range from the 18th most Republican to the 26th most Democratic districts into a separate plot, as these are the districts with the potential to be competitive.

This figure is somewhat striking; for the districts in the outer regions of the potentially competitive range, neither of the first two Staff Plans comes close to achieving competitiveness for these districts, despite significant portions of all three ensembles containing competitive districts in these positions. This data lends additional credence to our supposition that these Staff Plans may have incorporated priorities not included in our ensemble models, especially in light of the extremely competitive Second Staff Plan for the State Senate. We would welcome additional information from the Commission that would help us to better understand their priorities and to build more accurate ensembles for State House plans.

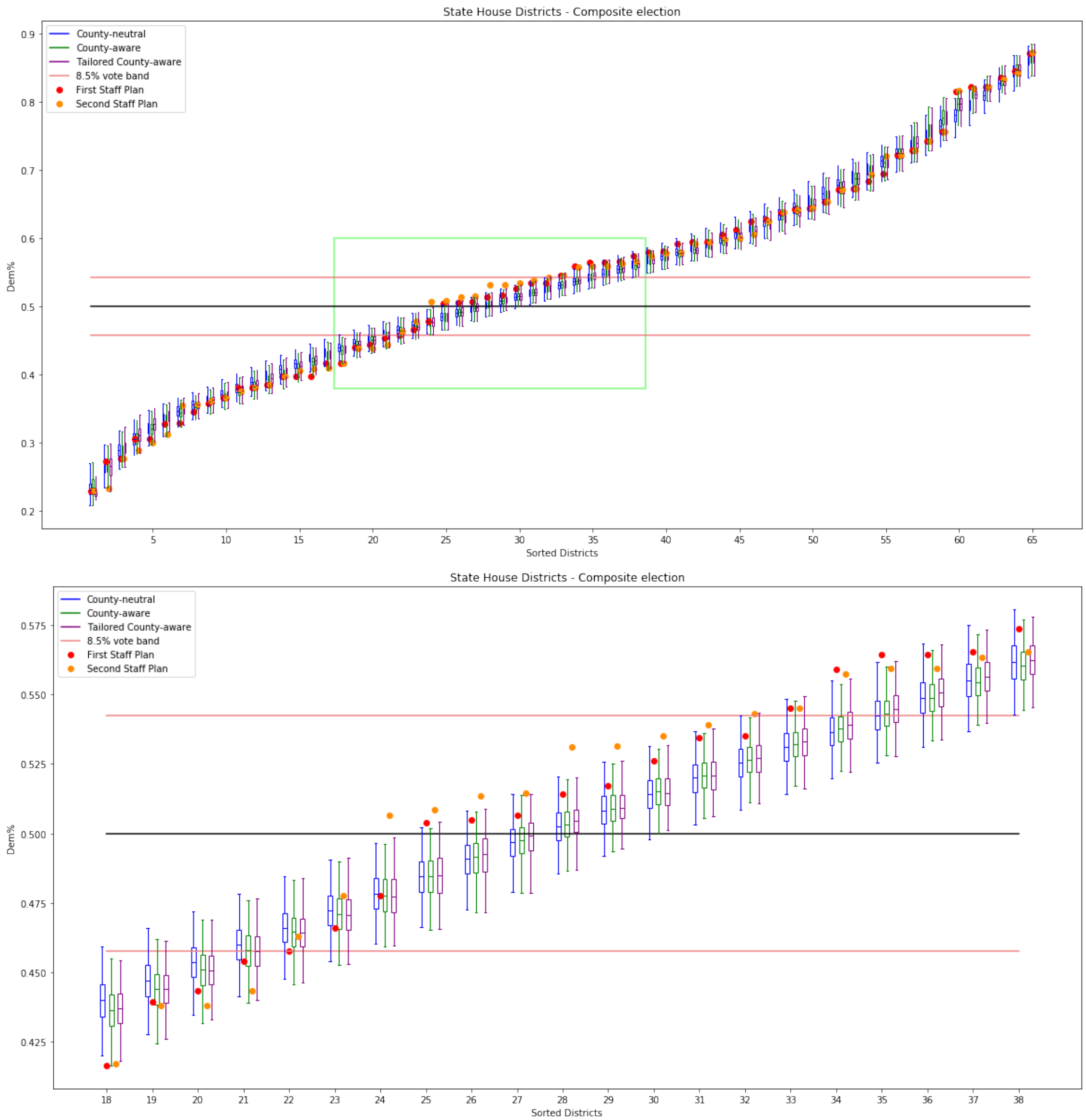


Figure 22: Democratic vote shares by district for ensembles and Staff Plans for State House, with competitiveness vote band

6.3 Partisan seat share

Finally, we compare the First and Second Staff Plans to our ensembles regarding partisan seat share. In Figure 23, we have added the values for the districts in both Staff Plans for the number of Democratic seats in the composite election to the histograms from Figure 18.

The First Staff Plan produces 41 Democratic seats, which is about 3 seats above the means of all three ensembles. This plan might reasonably be considered an outlier, although not an extreme

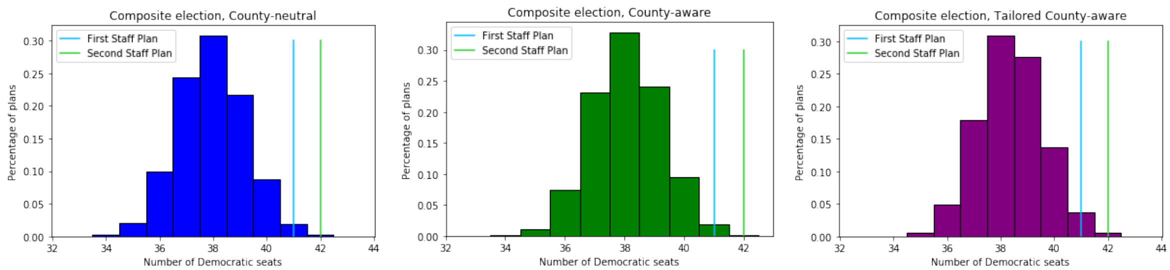


Figure 23: Numbers of Democratic seats won in ensembles and Staff Plans for State House

one. The Second Staff Plan produces 42 Democratic seats, which is about 4 seats above the means of all three ensembles and is definitely an extreme outlier; only 0.24% of plans in the county-neutral ensemble, 0.20% of plans in the county-aware ensemble, and 0.59% of plans in the tailored county-aware ensemble produce 42 or more Democratic seats.

The extremity of this plan is somewhat mitigated by the additional context contained in Figure 22; the bottom 4 Democratic seats in the Second Staff Plan are very competitive, with Democratic vote shares of 50.65%, 50.85%, 51.35%, and 51.45%. So these seats might reasonably be viewed as "toss-up" seats rather than either Democratic or Republican.

In order to explore this idea of "toss-up seats" further, we considered the range of partisan outcomes for the 8 statewide elections included in the composite election. Statewide Democratic vote shares for these elections ranged from 52.7% (President 2016) to 55.2% (Governor 2018), with an average for the composite election of 54.0%. This means that across these 8 elections, Democratic vote shares for the average district ranged from 1.3% below to 1.2% above the figure reported for the composite election. In particular, a typical district with reported Democratic vote share between 48.7% and 51.2% probably experienced majority votes for both parties at some point during these 8 elections.

With this in mind, we decided to explore an alternative classification of district-based partisan outcomes into three categories based on a 3% vote band about 50%:

1. Democratic: Democratic vote share of 51.5% or more;
2. Republican: Democratic vote share of 48.5% or less;
3. Toss-up: Democratic vote share between 48.5% and 51.5%.

The histograms in Figure 24 describe what percentage of plans in each ensemble fall into each of these three categories, with the values for the First and Second Staff Plans included for comparison.

With the addition of the toss-up category, the Staff Plans look somewhat less extreme relative to the ensembles. The First Staff Plan is within one or two seats of the most common outcomes for

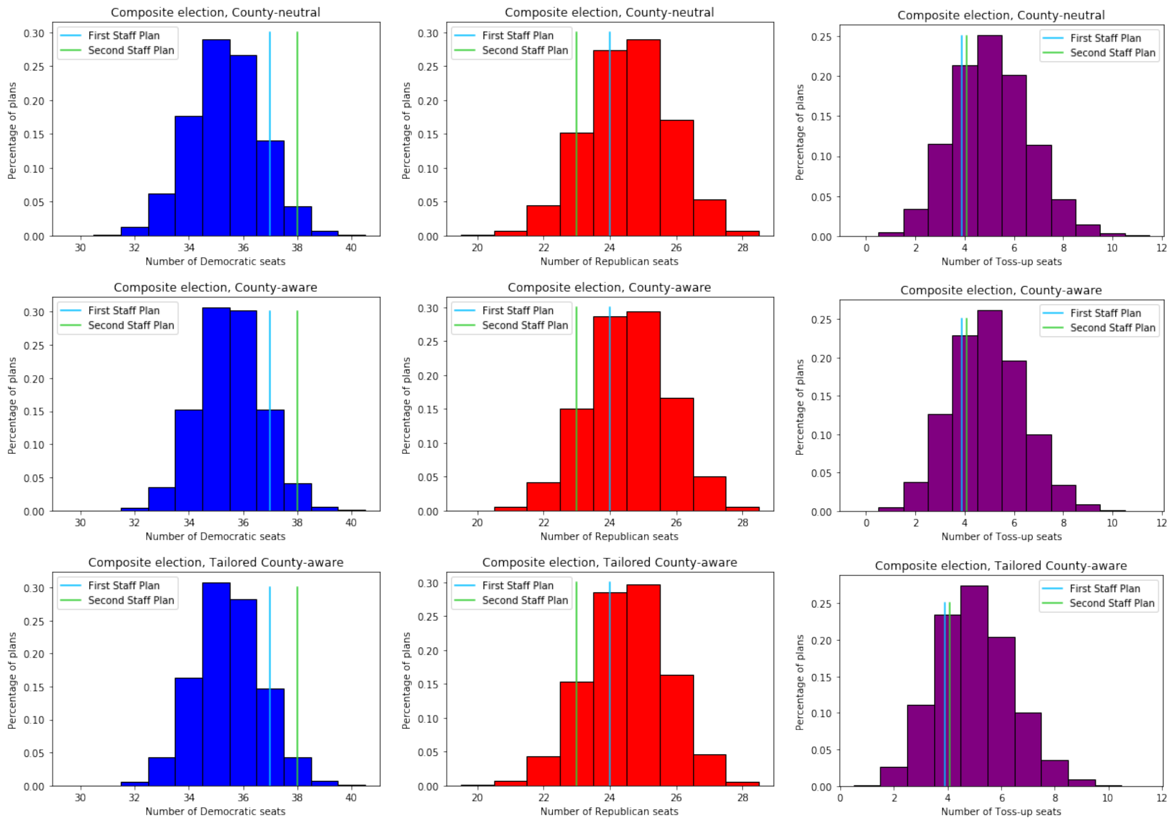


Figure 24: Numbers of Democratic, Republican, and Toss-up seats in ensembles and Staff Plans for State House

all ensembles with respect to the numbers Democratic, Republican, and toss-up seats. The Second Staff Plan is still something of an outlier with respect to the number of Democratic seats, but not the extreme outlier that it appears to be in Figure 23. The number of toss-up seats is one seat below the mean of all ensembles in both Staff Plans, which is consistent with the pattern for competitive seats within the 8.5% threshold for these plans.

6.4 Conclusions

The Commission and the nonpartisan staff have clearly put much thought and effort into the design of the First and Second Staff Plans for both the Senate and the House. Our computer-generated ensembles of plans cannot take into account the myriad of considerations that went into their design, or those that the Commission will prioritize for additional plans.

For the Senate, we do do not detect any evidence of problematic features in either Staff Plan. Both plans perform particularly well with respect to the numbers of districts with Hispanic voting age population above 30%, and the Second Staff Plan does an extremely good job of maximizing the number of competitive districts. Both plans are also within the ranges of reasonable outcomes for Democratic and Republican seat shares with respect to all three ensembles.

For the House, both Staff Plans perform well with respect to the numbers of districts with Hispanic voting age population above 30%, with the Second Staff Plan performing extremely well with 14 such districts, compared to 11 in the First Staff Plan. We have concerns about the numbers of competitive districts produced by both Staff Plans, which are within the range of reasonable outcomes but do not show evidence that the number of competitive districts has been maximized. We also have concerns that the number of Democratic seats produced by the Second Staff Plan is abnormally high, although this concern is somewhat tempered by the further exploration of which seats might reasonably be considered toss-ups. These effects may well be related to the relatively large number of districts with a high Hispanic voting age population, to the Commission's efforts to preserve other communities of interest, or to other priorities not included in our ensemble models, but we do not currently have enough information about such priorities to assess their effects.

7 Acknowledgments

We are very grateful to Moon Duchin and the MGGG Redistricting Lab for their pioneering work in ensemble analysis, for their open source python package GerryChain (available at https://github.com/mggg/GerryChain) which we have used for our computations, and for their ongoing support for our work.

References

[1] *League of Women Voters of Pennsylvania et al. v. Commonwealth of Pennsylvania et al.*, Supreme Court of Pennsylvania, 2018.

[2] *Common Cause v. Lewis*, Wake County, NC Superior Court, 2019.

[3] *Rucho v. Common Cause*, 588 U.S., 2019.

[4] Jeanne Clelland, Nicholas Bossenbroek, Thomas Heckmaster, Adam Nelson, Peter Rock, and Jade VanAusdall, *Compactness statistics for spanning tree recombination*, arXiv e-prints (2021), arXiv:2103.02699.

[5] Jeanne Clelland, Haley Colgate, Daryl DeFord, Beth Malmskog, and Flavia Sancier-Barbosa, *Colorado in Context: Congressional Redistricting and Competing Fairness Criteria in Colorado*, J Comput Soc Sc (2021), `https://doi.org/10.1007/s42001-021-00119-7`.

[6] Daryl DeFord, Moon Duchin, and Justin Solomon, *Recombination: A family of Markov chains for redistricting*, arXiv e-prints (2019), arXiv:1911.05725.

[7] Persi Diaconis, *The Markov chain Monte Carlo revolution*, Bull. Amer. Math. Soc. (N.S.) **46** (2009), no. 2, 179–205.

[8] Moon Duchin, *Outlier analysis for Pennsylvania congressional redistricting*, Expert report to Pennsylvania Governor Tom Wolf for League of Women Voters v. Commonwealth of Pennsylvania, 2018, accessed May 2020 at https://www.governor.pa.gov/wp-content/uploads/2018/02/md-report.pdf.

[9] Jonathan Mattingly, *Report on Redistricting: Drawing the Line*, Expert report to U.S. District Court for the Middle District of North Carolina for Common Cause v. Rucho, 2017, available at https://s10294.pcdn.co/wp-content/uploads/2016/05/Expert-Report-of-Jonathan-Mattingly.pdf.

[10] ———, *Expert Report on the North Carolina State Legislature*, Tech. report, 2019, available at https:// sites.duke.edu/quantifyinggerrymandering/files/2019/09/Report.pdf.

[11] Wesley Pegden, *Pennsylvania's Congressional districting is an outlier: Expert Report*, Expert report of Wesley Pegden in League of Women Voters of Pennsylvania v. Commonwealth of Pennsylvania, 2017, available at https://www.pubintlaw.org/wp-content/uploads/2017/06/Expert-Report-Wesley-Pegden.pdf.

[12] ———, Expert report of Wesley Pegden in Common Cause v. Lewis, 2019, available at https://www.commoncause.org/north-carolina/wp-content/uploads/sites/22/2019/09/Common-Cause-v.-Lewis-trial-court-decision-9.3.19.pdf, pp. 94-105.

A Technical details

A.1 Data collection

In order to build the precinct map used to generate ensembles, we obtained data from the following sources:

- A shapefile with the geographic boundaries of all 2020 voting precincts in Colorado, including precinct-level election results for all statewide elections in 2020, was given to us by Louis Pino from the Commission's nonpartisan staff.
- In the summer of 2019, the third author's student Haley Colgate compiled a shapefile with the geographic boundaries of all 2018 voting precincts in Colorado, including precinct-level election results for all statewide elections in 2018, with the assistance of Todd Bleess of the Colorado State Demography Office.
- A shapefile with the geographic boundaries of all 2016 voting precincts in Colorado, including

precinct-level election results for all statewide elections in 2016, was obtained from the Voting and Election Science Team's repository on the Harvard Dataverse at https://dataverse.harvard.edu/dataverse/electionscience.

- Population data from the 2020 Census was taken from the 2020 PL 94-171 Data Summary File for Colorado based on the Decennial Census at the Census Block Level, obtained from the Redistricting Data Hub at https://redistrictingdatahub.org.

The open source python package Maup, developed by the MGGG Redistricting Lab and available at https://github.com/mggg/maup, was used to aggregate/disaggregate all population and election data from their original geographies onto the precinct geographies in the 2020 precinct shapefile. The resulting shapefile contains all the data required to compute population and election results for any district composed of 2020 precincts.

For our tailored county-aware ensemble, we modified this shapefile as follows: For each of the 27 counties with 2020 Census population less than 10,000 and each of the 4 communities of interest identified by the Commission, we merged all precincts for that county or community of interest to create a single unit in the shapefile. Since districts in our ensemble plans are built from units in the shapefile, this guarantees that all plans in the tailored county-aware ensemble keep each of these counties and communities of interest whole within districts.

A.2 Ensemble generation

In order to generate our ensembles, we used the Recombination ("ReCom") method developed by the MGGG Redistricting Lab in 2018. (See [6] for a thorough treatment of this method.) For this method, the precinct map is modeled by a mathematical object called a **dual graph**, where each precinct is represented by a point called a **vertex**, and two vertices are connected by an **edge** if the precincts that they represent share a geographic boundary of positive length. A map of Colorado's 2020 voting precincts and its dual graph are shown in Figure 25.

A district plan is then represented by a partition of the dual graph into connected subgraphs, one for each district. (As an illustration, Figure 26 shows the graph partition corresponding to the First Staff Plan for Congressional districts.) A partition is **valid** if it represents a legally valid district plan; at a minimum, the districts in the plan should be contiguous and have (approximately) equal population.

An ensemble starts with one randomly constructed valid plan, called the "seed plan." The ensemble is then constructed by a mathematical process called a **Markov chain,** in which each new plan is created by applying a random process to modify the previous plan in some way. For the ReCom method used to build our ensembles, this random process works as follows: At each step, the algorithm randomly selects a pair of adjacent districts and merges the two subgraphs corresponding

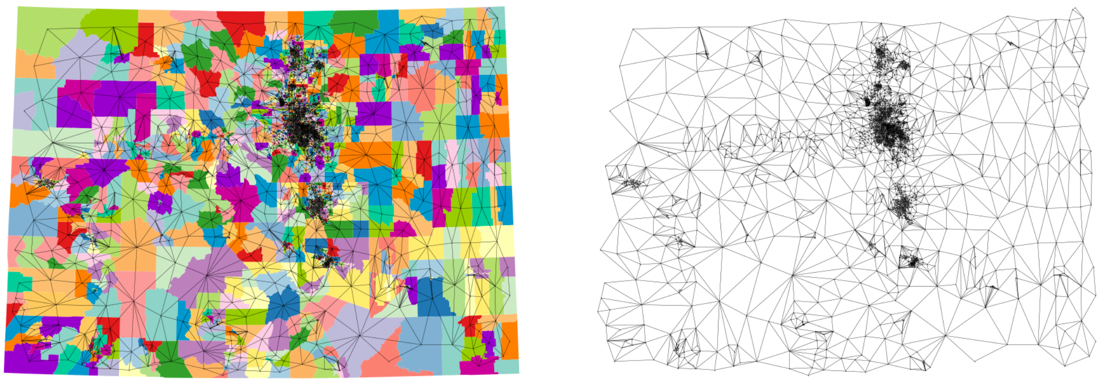


Figure 25: Colorado 2020 precinct map and dual graph

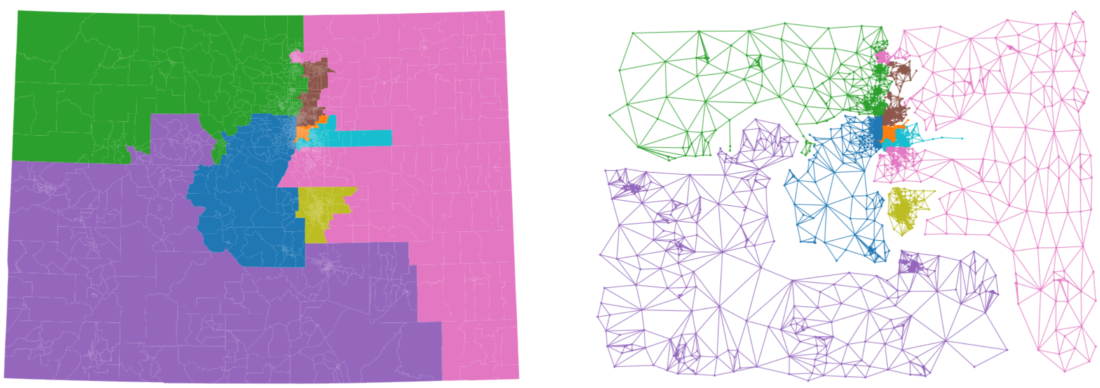


Figure 26: First Staff Plan for Congressional districts and corresponding dual subgraphs

to these districts into a single graph. Next, it generates a **spanning tree** for the merged graph—i.e., a subgraph consisting of all the graph's vertices and a subset of its edges, with the property that this subgraph is contiguous and has no closed loops—chosen randomly and uniformly from the set of all spanning trees of the merged graph. Finally, it looks for an edge to cut in order to create two new districts that each satisfy the population constraint. (District contiguity is automatic with this method.) This process is illustrated in Figure 27.

Part of the appeal of the Markov chain approach is a well-developed theory and a long history of applications of Markov chain sampling methods (see, e.g., [7]). In particular, a sufficiently long Markov chain is theoretically guaranteed to produce an ensemble that accurately represents a specific probability distribution on the entire space of valid district plans. In general, this probability distribution is difficult to determine explicitly, but for the ReCom method there is good heuristic and experimental evidence indicating that the probability of any particular plan appearing in the ensemble is closely related to a natural discrete measure for district compactness. In practice, this

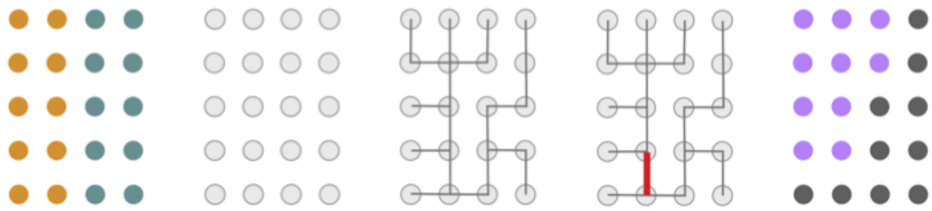


Figure 27: A ReCom step (Figure 4 in [6]; used with permission.)

means that this method is strongly biased towards plans with relatively compact districts and has no other detectable bias towards any particular type of plan (see, e.g., [4] and [6]). Our county-neutral ensemble was generated with this method; some examples of plans from this ensemble are shown in Figure 28.

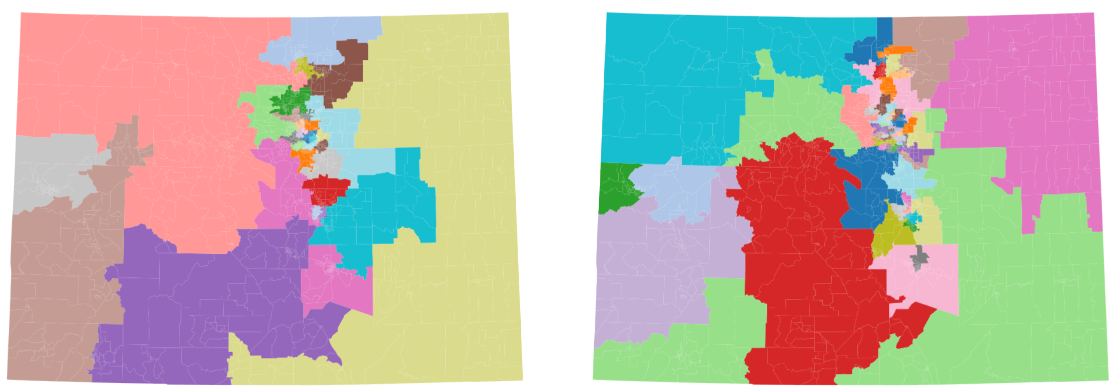


Figure 28: Examples of plans created by the ReCom method for county-neutral ensembles for Senate (left) and House (right)

For our county-aware ensemble, a variation was used in the construction of the spanning tree for the merged graph, in which the random choice of edges to form the spanning tree is more heavily weighted towards intra-county edges, so that the resulting spanning tree contains relatively few edges connecting precincts in different counties. When the tree is cut, it is less likely to produce districts that split counties. Some examples of plans from this ensemble are shown in Figure 29.

For our tailored county-aware ensemble, we applied the county-aware variation of ReCom to the shapefile obtained by merging precincts as described in Section A.1. Some examples of plans from this ensemble are shown in Figure 30.

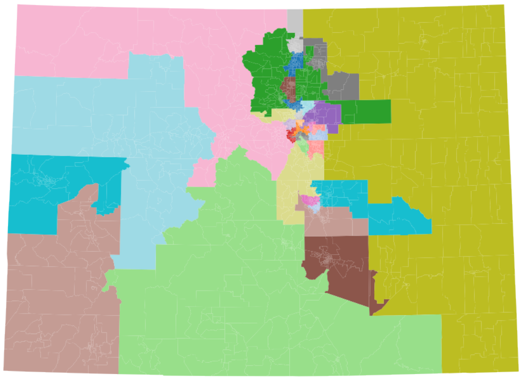
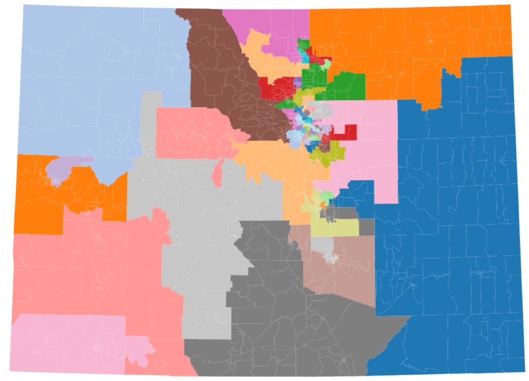
 

Figure 29: Examples of plans created by the ReCom method for county-aware ensembles for Senate (left) and House (right)

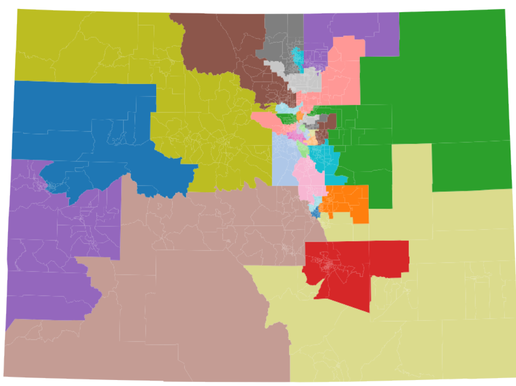
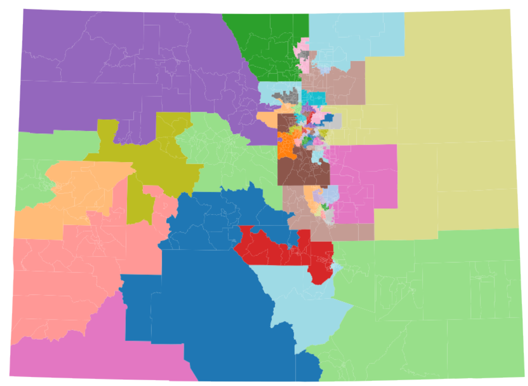
 

Figure 30: Examples of plans created by the ReCom method for tailored county-aware ensembles for Senate (left) and House (right)

A.3 Ensemble size

Regarding the question of how long is "sufficiently long" for a Markov chain to produce a representative sample of plans, there is unfortunately no good theoretical answer. This question is usually answered heuristically, by running chains until statistics of interest appear to stabilize in a way that is not dependent upon the choice of seed plan. This stabilization is referred to as "mixing" or "convergence" of the statistics being measured.

When two or more independently generated chains are available, convergence of relevant measures can be checked by comparing the distributions of the measures from the available chains. If the distributions are "close enough," we consider the chains to be sufficiently long for analyses. To check how close two empirical distributions are from one another, we used the two-sample Kolmogorov-Smirnov (KS) statistic, which is the maximum vertical distance between two empirical cumulative distribution functions (ECDFs) derived from two independent samples. KS values of 0.01 or lower are indicative of nearly indistinguishable chains from different seeds; however, this is a very high bar to meet and may require generating ensembles of several million steps.

When only one independently generated chain is available, it is still possible to check how fast the chain is mixing by looking at the chain's autocorrelations, which measure how dependent a chain is on its previous steps. Autocorrelations take values between -1 and 1, with a value of 0 indicating non-dependence, and values close to 1 or -1 indicating high dependence. For example, a lag 10 autocorrelation of 0.8 would indicate that any step of the chain is highly dependent on its 10 previous steps. For fast-converging chains, autocorrelations quickly decay to 0 as the lags increase. Inspecting which lags give autocorrelations close to 0 is a way of checking how quickly a chain is converging and it also allows one to make rough estimates on how long the chain should be for appropriate analyses. For instance, if the lag 1000 of a chain is close to 0, then we can consider each step of the chain to be fairly uncorrelated with 1000 steps prior. Therefore, if we were to collect every 1000th step of the chain, we would obtain a fairly uncorrelated sample, which would be close enough to an independent sample. This means that for a chain of size 2 million with a lag 1000 autocorrelation close to 0, we can expect the chain to be at least as good as an independent sample of size 2000, which would be long enough for reliable analyses and outlier detection.

Among the three types of chains used in this report (county-neutral, county-aware, and tailored county-aware), we explored convergence in more detail for county-aware chains for the State House, since State Senate chains are expected to have much faster convergence than House chains because an increased number of districts significantly slows down the rate of convergence. We also focused our convergence analyses on two types of measures: vote shares and seat shares. In our experience, other types of measures (for example, number of competitive districts, number of counties split, etc.) have convergence speeds similar to or better than vote shares and seat shares.

County-aware chains:

For the convergence analysis of county-aware chains for the State House, we used three races to represent the proportion of Democratic votes at the precinct level: 2018 Attorney General (AG18), 2020 President (PRES20), and 2018 Secretary of State (SOS18). For each of these three races, 7 chains of size 2 million were generated using 7 independent seeds (starting plans) and we calculated the two-sample KS statistic for each pair of seeds (21 pairs).

The three largest two-sample KS statistics were 0.2318, 0.2275, and 0.2273 for the vote shares of the least Democratic district for the SOS18, PRES20, and AG18 races respectively. Approximately 11% of the KS values for vote shares were below 0.01 and 11% were greater than 0.05. These values are indicative of slow convergence for most of the vote shares.

Autocorrelations also indicated slow-mixing chains. The minimum lag such that autocorrelations were less than 0.01 varied widely. For example, for the AG18 race data, the smallest such lag was 2056 and there were 6 lags above 200,000. If we were to require a sample at least as large as 500 times the largest of the minimum lags, we would need sample sizes greater than 100 million.

Even though our analysis indicates that chains of size 2 million for the State House do not have ideal convergence diagnostics, in order to detect outliers we only need to be able to trust the estimates for low and high quantiles for measures of interest generated by our chains. That is, we would like the lower and higher sample quantiles of measures generated by different seeds to be "close enough" so that we can trust the detection of unusual observations even when chains are not very well mixed.

To explore how long our chains should be for outlier detection, we looked at the differences between the tails of the ECDFs of samples generated by different seeds. We considered the tail of a distribution to be the top and bottom 2.5% of the distribution, that is, all observations beyond the middle 95% of the distribution. For chains of size 2 million, the largest two-sample tail distance was 0.044 for the second least democratic district. Approximately 87% of the tail distance values for vote shares were below 0.01. This number was 73% for seat shares. The largest difference between the tail quantiles was 0.018 for the 56th least democratic district. Approximately 95% of the differences between the tail quantiles were below 0.01.

These summaries indicate that a sample of size 2 million is able to effectively identify unusual observations.

County-neutral chains:

County-neutral chains usually have faster convergence than county-aware chains; here we confirmed this by looking at their autocorrelations. Therefore, we are fairly confident that county-neutral chains are at least as appropriate as county-aware chains for analyses and detection of unusual observations.

Tailored county-aware chains:

Tailored county-aware chains showed slower convergence than non-tailored county-aware chains, and there were more districts with a KS statistic above 0.2 than there were in non-tailored chains. To mitigate this and improve the accuracy of estimates produced by tailored samples, we constructed a chain of size 2 million by combining every other step of two independently generated chains of size 2 million. The expectation is that two independently generated chains will together cover a larger portion of the sample space of plans than one chain alone. Therefore, this combination is expected to give a better representative sample, closer in accuracy to the non-tailored county-aware chains. We also note that all statistics produced by the tailored chains (except for counties split and total county splits, as expected) were very similar to those produced by the untailored chains, so no concerns were raised that might be attributable to inadequate sample size.